

THÈSE

présentée

DEVANT L'UNIVERSITÉ CLAUDE - BERNARD DE LYON

pour obtenir

le diplôme de Docteur de Spécialité (3° Cycle)
de Sciences Biologiques (Biologie Appliquée)

PAR

Jean JACQ

CONTRIBUTION A L'ÉTUDE DE LA FORMATION ET DE L'IDENTIFICATION DES CLASSES EN BIOLOGIE

soutenue le 27 Avril 1972

devant la Commission d'examen

MM. LEGAY

Président

TEULADE

LAZAR

DEFAYOLLE

} Examineurs



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41552613_0025

THÈSE

présentée

DEVANT L'UNIVERSITÉ CLAUDE - BERNARD DE LYON

pour obtenir

le diplôme de Docteur de Spécialité (3° Cycle)
de Sciences Biologiques (Biologie Appliquée)

PAR

Jean JACQ

CONTRIBUTION A L'ÉTUDE DE LA FORMATION ET DE L'IDENTIFICATION DES CLASSES EN BIOLOGIE

soutenue le 27 Avril 1972

devant la Commission d'examen

MM. LEGAY

Président

TEULADE

LAZAR

DEFAYOLLE

} Examineurs

UNIVERSITE CLAUDE BERNARD
LYON I

Président : M. le Professeur J. BOIDIN.

1er Vice-Président : M. R. TOURAINÉ, Maître de Conférences agrégé.

2ème Vice-Président : M. P. PONCET, Maître-Assistant.

3ème Vice-Président : M. D. SETTELEN, Etudiant.

UNITES D'ENSEIGNEMENT ET DE RECHERCHE

<u>UNITES</u>	<u>DIRECTEURS</u>
U.E.R. médicale Grance-Blanche	M. le Professeur D. GERMAIN
U.E.R. médicale Alexis-Carrel	M. le Professeur C. GIROD
U.E.R. médicale LYON-Nord	M. le Professeur A. BERTOYE
U.E.R. médicale LYON-Sud-Ouest	M. le Professeur L. TOLOT
U.E.R. des Sciences pharmaceutiques	M. M. CARRAZ, Maître de Conférences agrégé
U.E.R. des Techniques de Réadaptation	M. le Professeur P. MOUNIER-KUHN
U.E.R. de Biologie humaine	M. J.C. CZYBA, Maître de Conférences agrégé
U.E.R. d'Education Physique et Sportive	M. J. SAPIN, Professeur E.P.S.
U.E.R. des Sciences Odontologiques	M. R. VINCENT
U.E.R. de Mathématiques	M. le Doyen J. BRACONNIER
U.E.R. de Physique	M. le Professeur M. DUFAY
U.E.R. de Chimie et Biochimie	M. le Professeur D. GAUTHERON
U.E.R. des Sciences de la Nature	M. le Professeur L. DAVID
U.E.R. de Biodynamique et Psychopédagogie	M. le Professeur R. FONTANGES
U.E.R. de Physique Nucléaire	M. le Professeur A. SARAZIN
U.E.R. de Mécanique	M. le Professeur J. MATHIEU
Observatoire	M. le Professeur J.H. BIGAY
Institut Universitaire de Technologie I	M. le Professeur B. POUYET
Institut Universitaire de Technologie II	M. J. GALLET

Secrétaire Général : M. P. IZAUTE.

Les recherches qui font l'objet de ce travail ont été menées sous la direction de Monsieur le Professeur LEGAY. Il a bien voulu nous laisser une entière liberté dans la conduite de nos travaux tout en nous faisant bénéficier de ses conseils. C'est à lui que nous devons notre attitude relative à la notion de modèle.

Depuis plusieurs années, nous avons le plaisir de travailler sous la direction de Monsieur le Médecin en Chef DEFAYOLLE ; il a su nous former aux techniques de l'analyse multidimensionnelle, nous guider et nous encourager dans la réalisation de ces études. Qu'il trouve ici l'expression de notre profonde gratitude.

Nous sommes particulièrement reconnaissant à Monsieur le Professeur LAZAR d'avoir accepté de faire partie de ce jury ; nous n'oublions pas qu'il nous a initié à la méthodologie statistique dans le cadre du Centre d'Enseignement des Statistiques Appliquées à la Médecine et à la Biologie.

Nous remercions Monsieur TEULADE, Maître de Conférences, de l'attention qu'il a bien voulu nous témoigner en acceptant de juger ce travail.

Ces études effectuées à la Division de Psychologie n'ont pu l'être que grâce à la compréhension et à l'aide matérielle apportée par la Direction du Centre de Recherches du Service de Santé des Armées.

Nos plus vifs remerciements vont à Madame PAVÉ et au personnel du Centre de Recherches du Service de Santé des Armées à Lyon qui ont grandement contribué à la réalisation de cet ouvrage.

S O M M A I R E

INTRODUCTION	1
I. LA FORMATION ET L'IDENTIFICATION DES CLASSES DANS LES SYSTEMES BIOLOGIQUES ET ARTIFICIELS	3
ASPECTS FONDAMENTAUX	3
A. LES SYSTEMES BIOLOGIQUES NATURELS	4
B. LES SYSTEMES ELABORES PAR LES SOCIETES	5
C. LES SYSTEMES AUTOMATISES	6
LES MEMOIRES	7
LA STRUCTURE DES INFORMATIONS STOCKEES	8
II. ELABORATION DU MODELE	10
A. HYPOTHESES ET ANALOGIES FONDAMENTALES	10
1) Description des objets	10
2) Analogie géométrique	11
3) Relations d'analogies	11
B. ELABORATION DU PROCESSUS	15
1) Processus de classification	16
2) Processus d'identification	18
3) Processus de structuration - regroupement	18
C. PRESENTATION DU PROGRAMME	20
1) Présentation des données	20
2) Présentation des résultats	22
D. ESSAIS EFFECTUES	23
Essai n° 1 : Utilisation de la distance euclidienne	23
Essai n° 2 : Utilisation de la métrique hypersphérique	33
Essai n° 3 : Utilisation des caractères qualitatifs	33

III. APPLICATIONS	35
A. APPLICATIONS RELEVANT DE LA MICROBIOLOGIE	35
- Description de la microméthode	37
1) Travaux effectués sur les Lactobacilles	37
2) Travaux effectués sur les Klebsiellae	37
3) Travaux effectués sur les levures	41
- L'aspect système	43
- Conclusions concernant les applications microbiologiques	43
B. APPLICATIONS RELEVANT DE LA BIOLOGIE CLINIQUE	45
- Echantillon étudié	45
- Etudes effectuées	45
1) Etudes sur les notes en facteurs	46
2) Etudes sur les variables standardisées	47
- Interprétation et conclusion	47
IV. DISCUSSION GENERALE ET CONCLUSION	48
- Quelques aspects des classes homogènes	48
1) Le modèle centroïde	50
a) Variables quantitatives	50
b) Variables qualitatives	52
2) Le modèle à enchaînement	54
- Choix de la métrique	54
- Résultats obtenus en fonction de la nature des classes	55
- Conclusion	55
BIBLIOGRAPHIE	57
PROGRAMME CLASSID	I à XII

I N T R O D U C T I O N

Le travail dont nous allons faire l'exposé se situe dans le cadre d'une recherche générale entreprise en vue de dégager une voie d'approche des problèmes relatifs à l'identification. Sous des aspects très divers de nombreuses études ont été et sont encore fréquemment effectuées sur ces sujets. Au cours des décennies écoulées les psychologues se sont longuement préoccupés des divers aspects de la perception ; plus récemment, différents modèles ont été proposés en vue d'apporter une solution à toute une catégorie d'études désignées sous le terme général de "reconnaissance de forme". Par ailleurs, des moyens de traitement de l'information particulièrement importants rendent possible l'application de certains processus qui autrefois n'auraient eu qu'un caractère très spéculatif.

Examinons comment s'effectue cette identification : il est rare qu'un objet soit entièrement nouveau ; en fait il ressemble souvent plus ou moins à une catégorie d'objets que l'on a déjà rencontrés. Pour reconnaître cet objet, il faut donc le classer par rapport aux objets déjà rencontrés. S'il n'est pas possible de déterminer actuellement comment se sont effectuées ces premières classifications ; il nous a semblé qu'une étude, même sommaire, sur la formation des classes dans les systèmes biologiques et artificiels était susceptible d'aider à la formulation de ces problèmes.

Dans ce travail, le terme "classification" concernera plus particulièrement le fait de former des classes à partir d'un ensemble d'objets non structurés ; tandis que le terme d'"identification" sera appliqué au processus qui consiste à assigner un objet à des classes préalablement définies. Il peut paraître arbitraire de distinguer ainsi deux mécanismes qui sont intimement imbriqués dans la nature, mais il importe de remarquer qu'ils sont souvent étudiés séparément et cela sans que les relations qui les lient soient même évoquées. L'utilisation de méthodes relevant de la taxonomie pour approcher les problèmes d'identification est une des idées fondamentales de ce travail (JARDINE, 1971). En fait, le but des recherches de ce type est souvent d'obtenir une méthode d'identification, mais nous avons remarqué qu'il était difficile de définir les classes auxquelles devaient être assignés les ob-

jets. Cette constatation est à l'origine de la démarche qui a été suivie. Certes toutes les méthodes de classification ne pourraient pas être appliquées à des fins d'identification et cela pour différentes raisons : certaines, concernant le traitement de l'information et seront évoquées dans la deuxième partie lors de la formulation du modèle ; d'autres relèvent du problème plus fondamental de la notion de classe et seront abordés dans les applications de la troisième partie ainsi que dans la discussion générale.

Dans ce contexte, les termes définis dans le paragraphe précédent, ainsi que ceux dont les équivalences sont données ci-dessous, ne visent en aucune manière à fixer une terminologie très diversifiée en fonction des domaines concernés ; encore moins à présumer l'existence dans la nature des différences qu'elles pourraient impliquer, mais uniquement à rendre l'exposé plus aisé et plus commode la transition entre les documents rappelés en référence.

PRINCIPALES EQUIVALENCES

Classification
Taxonomie
Typologie
"Cluster analysis"

Identification
Reconnaissance de forme
"Pattern recognition"

Classe
Taxa
Groupe
Ensemble
• Configuration
"Cluster"

Objet
Elément
Individu

I. LA FORMATION ET L'IDENTIFICATION DES CLASSES DANS LES SYSTEMES BIOLOGIQUES ET ARTIFICIELS

Il semble bien que les nécessités de la vie de relation jointes à celles de la perpétuation de l'espèce aient contribué à développer, chez les animaux, des aptitudes à élaborer des classifications d'êtres nécessaires à leur survie en tant qu'individu ou espèce. Ces processus s'effectuent dans le cadre d'un système qui comprend outre "l'organisme identifiant" l'environnement dans lequel il vit. Les identifications effectuées par les animaux ainsi que par les humains (dans certains cas) peuvent être considérées comme appartenant à cette catégorie.

Lorsqu'il s'est agi de classer avec une discrimination plus grande les espèces biologiques en vue d'études de caractère scientifique ; les chercheurs ont, au cours des siècles, élaboré des procédés plus méthodiques qui ont permis de mettre au point la plupart des classifications systématiques encore en usage aujourd'hui. Cette deuxième catégorie se différencie de la première par l'aspect plus ou moins scientifique de la méthode suivie et l'aspect plus ou moins exhaustif des échantillons étudiés. A titre d'exemple, on peut considérer que les identifications de maladies ou de microbes appartiennent à cette catégorie.

Les recherches entreprises en vue de transformer dans le sens d'une mécanisation plus ou moins complète les systèmes de la deuxième catégorie, ainsi que certaines de leurs difficultés sont à l'origine des études relatives aux systèmes automatisés qui constituent la troisième catégorie. L'étude des systèmes biologiques en vue de réaliser une machine présentant de telles performances représente l'aspect spécifiquement bionique ; par ailleurs, ces recherches entraîneront l'élaboration de modèles qui constitueront parfois des outils intéressants pour l'étude des phénomènes biologiques eux-mêmes ; ce qui correspond à l'aspect cybernétique.

ASPECTS FONDAMENTAUX

L'aspect le plus important des processus d'identification est certainement le fait qu'il s'agit d'un traitement d'information :

"Information is information
no matter no energy"

(N. WIENER, 1948)

C'est ce qui permet de penser que ces processus se retrouvent exclusivement dans le règne animal ; car il semble bien que les végétaux se bornent, pour la vie de relation, à un traitement de matière qu'ils prélèvent dans le milieu où ils se trouvent et pour la reproduction : soit à des modes asexués, soit à des modes sexués mais à forte probabilité de rencontre. Certes, l'objectif est bien - du moins en ce qui concerne la vie de relation dans ses aspects les plus primitifs - d'aboutir à un traitement de matière ne fusse que pour s'alimenter, mais ce traitement est précédé chez les animaux par une phase d'identification. De plus l'identification ne serait d'aucune utilité si "l'organisme identifiant" ne pouvait réagir sur le milieu pour arriver à ses fins ; d'où l'aspect stimulus-réponse de ces systèmes, aspect qui est lié aux possibilités de déplacement des animaux.

Le deuxième aspect est que si une espèce animale procède à des identifications à un stade donné de son développement c'est qu'il y a eu, au cours de la vie de l'individu ou au cours de l'évolution de l'espèce, une phase de formation de classe. Cela implique que l'espèce a reçu ou a élaboré un **programme** de traitement des informations relatives à ces classes et que, par ailleurs, ces informations ont été stockées dans une mémoire.

Enfin, chez la plupart des organismes les phases de formation de classe et d'identification correspondent à des périodes différentes : en principe l'organisme apprend au début de sa vie et identifie surtout ensuite.

A. LES SYSTEMES BIOLOGIQUES NATURELS

Les espèces peu évoluées effectuent des identifications : soit dans le cadre de la vie de relation dont le but est d'assurer la survie de l'individu, soit dans la vie sexuelle dont la finalité est la conservation de l'espèce. Pour la vie de relation, l'essentiel est de "manger et de ne pas être mangé" cela entraîne la reconnaissance de proies ou de prédateurs et de sites écologiques favorables ou défavorables : l'ensemble se traduisant par les réactions de poursuite ou de fuite. Les nécessités liées à la perpétuation de l'espèce entraînent la reconnaissance d'animaux de même espèce et l'identification du sexe (dimorphisme sexuel). Ainsi, on constate qu'il suffit d'un nombre relativement restreint de classes pour assurer ce traitement des données extérieures.

A ce nombre restreint de classes correspond également un petit nombre de variables en rapport avec celui des classes. Le traitement de l'information est également peu important : il nécessite uniquement des combinaisons relativement simples de ces variables ou s'effectue même parfois avec des variables spécifiques à chaque classe. En général, ces variables sont du type nominal (bleu, vert...), ou du type ordinal (gros, moyen, petit). Ce sont souvent des variables de perception visuelle, olfactive ou auditive. La formation des classes met en jeu une mémoire génétique spécifique de l'espèce ou un apprentissage sociologique dans les espèces les plus évoluées (TINBERGEN, 1967 - RICARD, 1971).

Si un tel système est suffisant malgré ses imperfections, puisqu'il a permis à certaines espèces de survivre pendant des ères géologiques, cela tient au fait que la nature compense son manque de performance par des processus qui assurent une plus grande probabilité de rencontre ou de dissémination de l'espèce. Ce phénomène se manifeste de plus en plus lorsqu'on descend dans l'échelle animale où on retrouve, avec les animaux les plus primitifs, des caractéristiques qui se rapprochent de celles des végétaux.

Lorsque les espèces deviennent plus évoluées, ce qui est le cas des animaux supérieurs et plus particulièrement de l'Homme, ces processus deviennent insuffisants tant pour les besoins quotidiens que pour ceux qui relèvent de la technologie ou des sciences.

B. LES SYSTEMES ELABORES PAR LES SOCIETES

Les processus mis en oeuvre dans ces systèmes résultent d'une lente élaboration sociale plus ou moins méthodique. "La classification populaire (objet d'étude de l'éthnobiologie) obéit à des règles génératrices aussi complexes, quoique non codifiées, que les classifications scientifiques" (ANDERSON, 1971). Pour ces dernières la nécessité d'obtenir une performance meilleure sur le plan de la discrimination a conduit les hommes à essayer de décrire les animaux et les choses qui constituaient leur environnement par des observations plus minutieuses conduisant ainsi à un recueil plus important d'informations. Les variables ainsi obtenues ont fait intervenir des observations microscopiques et diverses mensurations, auxquelles sont venues se rajouter les dosages chimiques ou biologiques et des mesures physiques parfois très élaborées.

Ce surcroît d'informations nouvelles n'aurait abouti qu'à l'accumulation d'un fatras de données s'il ne s'était accompagné de procédures de traitement

adaptées. On doit à LINNÉ d'avoir, dès le 18^e siècle, jeté les fondements d'une méthode qui a permis d'aboutir aux classifications systématiques des sciences naturelles encore en usage aujourd'hui (THEODORIDES, 1965). Ce système est basé sur la construction d'un ensemble de classes qui s'incluent les unes dans les autres de manière à former une structure hiérarchique constituant un arbre de classification. La systématique botanique représente assez bien la manière dont un tel système est construit. Son principal intérêt est qu'il constitue (selon la terminologie de COUFFIGNAL) une théorie déductive empirique : en suivant les embranchements successifs de l'arbre, il est possible de déterminer la classe à laquelle appartient un individu et d'en déduire ainsi ses autres propriétés.

Ce système, basé sur des dichotomies successives, est particulièrement bien adapté au traitement des variables qualitatives qui correspondent aux grands embranchements des règnes végétaux et animaux. Par contre, le traitement des variables quantitatives n'est pas possible sans perdre une quantité importante de leur information, ce qui diminue singulièrement les performances et l'intérêt de ce système lorsqu'on descend vers les rameaux constitués par les genres et les espèces.

Par ailleurs, au niveau de certains phylums la présence ou l'absence d'un caractère peut également être aléatoire, son traitement est alors impossible.

Les insuffisances de ce mode d'exploitation, notamment pour les variables quantitatives, sont en partie à l'origine des recherches entreprises actuellement.

C. LES SYSTEMES AUTOMATISES

S'il n'est pas possible de dégager de manière exhaustive les divers aspects de ces systèmes, qui pour la plupart n'en sont encore qu'à un stade conceptuel, on peut néanmoins essayer de déterminer les tendances de ces recherches et les performances qu'on souhaiterait obtenir.

Les études entreprises dans ce domaine ont eu pour objet de faire effectuer par une machine certaines tâches fastidieuses. C'est ainsi qu'on s'est intéressé à la lecture automatique des caractères alphanumériques en vue de faciliter le tri postal ou le transfert des informations à une machine. Dans les recherches de ce type l'ensemble des classes à identifier est préalablement défini ; mais ce n'est pas toujours le cas. Même s'il existe une classification, les utilisateurs estiment souvent que celle-ci doit être constamment adaptée pour tenir compte des informations nouvelles que les procédures

d'acquisition des données, sans cesse améliorées, permettent d'obtenir, et des apports d'éléments nouveaux aux ensembles précédemment étudiés.

Ces recherches peuvent donc concerner uniquement l'identification d'un ensemble de classes préalablement définies ou la formation des classes à partir d'un ensemble d'éléments en vue d'une identification ultérieure. Ces éléments peuvent être des objets ou des êtres vivants, mais peuvent aussi comprendre des concepts ou des notions, voire des contextes, des situations ; on emploiera dans ces derniers cas plus fréquemment le terme "reconnaissance de configuration" (GERARDIN, 1968). Enfin, le programme peut être élaboré par l'organisme en relation avec le milieu par une série d'essais successifs sanctionnés par des réponses (BELIS, 1966).

Ces trois catégories :

- identification de classes préalablement définies
- formation des classes en vue de leur identification
- élaboration d'un programme de traitement par essais successifs

ont en commun certaines caractéristiques que nous examineront brièvement avant d'aborder la formulation d'un modèle de la deuxième catégorie qui correspond à notre étude.

LES MEMOIRES

Les mémoires utilisées dans les différents systèmes peuvent être caractérisées par leur délai d'accès. Pour l'homme divers travaux ont mis en évidence l'existence d'une mémoire à court-terme et d'une mémoire à long-terme (ATKINSON, 1969 - WRIGHT, 1970).

La mémoire à court-terme est une mémoire de travail : la persistance des informations n'y dépasse guère trente secondes, à moins qu'elles ne soient constamment remises en circulation par un mécanisme particulier. Dans certaines conditions, ces informations peuvent passer dans la mémoire à long-terme pour y être stockées.

Ces deux mémoires permettent à l'individu de traiter les données qu'il reçoit ; mais dans les sociétés l'individu reçoit au cours de sa vie une certaine quantité d'informations qui est le fruit de l'élaboration collective des individus qui l'ont précédé. Il s'agit d'une véritable "mémoire socio-culturelle" qui est détenue par les membres de cette société. D'autres formes de stockage, tels que les livres, et actuellement tous les supports usuels de l'information, sont également utilisées. La transmission est effectuée

par l'observation des adultes ou par l'apprentissage.

Au niveau de l'espèce, la transmission de l'information est réalisée par une mémoire génétique. Au sens usuel du terme on entend par information génétique celle qui concerne des caractères tels que la forme du nez ou des oreilles ; mais il semble bien que, dans certains cas, elle puisse également contenir des programmes de comportement tels que ceux qui déterminent les conduites instinctuelles. D'après divers travaux, il ne serait pas exclu que ces mémoires contiennent aussi des informations relatives à certaines classes d'objets ; ce qui impliquerait des communications entre la mémoire à long-terme et la mémoire génétique (voir à ce sujet les travaux de SAUER relatés par RICARD (1971), dans son livre sur les migrations animales).

Le terme de mémoires biologiques s'applique aux mémoires à court-terme, à long-terme et aux mémoires génétiques. La nature chimique de ces dernières est aujourd'hui établie et il semble bien que les mémoires à long-terme soient aussi à base chimique.

Les mémoires des ordinateurs, qui sont utilisés dans certains systèmes automatisés, sont également classées en fonction de leur capacité et de leur temps d'accès. Les termes de mémoire centrale et de mémoire de masse (disques, bandes magnétiques) correspondent aux aspects qui ont été évoqués à propos des mémoires biologiques.

LA STRUCTURE DES INFORMATIONS STOCKEES

La localisation et la structuration des mémoires biologiques est certainement très différente de celles des mémoires utilisées dans les ordinateurs. Dans ces derniers, une information est stockée en un lieu bien déterminé et repérée par une adresse ; alors qu'il n'a pas été possible de déterminer la localisation des informations dans la mémoire à long-terme. "La chirurgie montre que, même dans des cas extrêmes où la maladie oblige à amputer un homme d'une partie notable de son cerveau, la mémoire reste pratiquement intacte" (GERARDIN, 1968).

La structure des informations stockées dans les mémoires biologiques est également mal connue et il semblerait que la démarche analogique soit à la base de la recherche des informations dans ces systèmes (GRAUFE et LYNN, 1969). Pour les systèmes relevant des sociétés humaines, nous avons vu que la structuration reposait sur la notion de classe. Comme la plupart des concepts fondamentaux, cette notion est assez difficile à définir (GRIZE, 1967) ;

pour COUFFIGNAL, "une classe est l'ensemble des êtres qui ont un même jeu d'attributs". Cette définition présente l'avantage de permettre la construction de différentes classes à partir d'opérations logiques telle l'intersection ou la réunion. Lorsqu'on aborde des problèmes tels que ceux qui concernent la reconnaissance des formes, il n'est pas toujours possible de définir un ensemble d'êtres par un même jeu d'attributs.

D'après la théorie des ensembles, un ensemble d'éléments peut être défini soit en extension (par l'énumération de tous les éléments constitutifs), soit en compréhension (à partir d'un certain nombre de propriétés). Une définition en extension ne peut être envisagée dans le cadre de ce travail, car elle correspondrait à une accumulation d'objets, sans formation de classe, ce qui aurait pour résultats de ne pas répondre au but cherché et d'encombrer, de manière excessive, la mémoire.

L'activité qui consiste à partir d'un ensemble d'êtres, à les classer en fonction de leur analogie et par rapport à un critère, relève du domaine de l'abstraction ; elle consiste à dégager les aspects généraux des êtres qui constituent une classe en négligeant les aspects spécifiques de chaque être. Suivant les critères et les analogies choisies, les classes obtenues peuvent être très différentes et on peut considérer qu'il existe plusieurs modèles de classe. Sans exclure l'intérêt que peuvent présenter d'autres modèles, il nous a semblé assez logique d'élaborer un modèle sur la base d'une relation de ressemblance, deux êtres appartiennent à une même classe si globalement ils se ressemblent. En fait, une classe peut être considérée comme représentée par un être moyen autour duquel sont dispersés les autres éléments. Ce type de classe "centroïde" sera développé et précisé ultérieurement en fonction des nécessités.

II. ELABORATION DU MODELE

Elaborer un modèle c'est choisir de conserver certaines propriétés de la réalité au détriment d'autres qui ne seront pas prises en considération. Les choix effectués ici tiennent compte des différents aspects dégagés au cours de la première partie et des contraintes qui résultent du traitement de l'information. Afin de rendre plus clair l'exposé, seules la formulation et la présentation seront abordées dans cette partie, les discussions relatives aux choix biologiques et aux conséquences qu'ils impliquent seront étudiées à propos des applications de la troisième partie ou dans la discussion générale.

A. HYPOTHESES ET ANALOGIES FONDAMENTALES

1) Description des objets

On suppose que l'échantillon à étudier comprend n objets, chaque objet étant décrit par k variables. L'ensemble des informations relatives à cet échantillon peut donc être représenté sous la forme d'une matrice rectangulaire à n lignes et k colonnes. Chaque ligne correspond à un objet et chaque colonne à une variable. S'il faut stocker la matrice, l'encombrement mémoire nécessaire M est :

$$M = n.k \quad \text{si } n = 1000 \text{ et } k = 50 \quad M = 50000$$

Si l'on se propose d'étudier les relations binaires sur l'ensemble des objets, il faudra calculer et stocker (au cas où on souhaite réutiliser les résultats) une matrice carrée symétrique de dimensions $n.k$, dans ce cas :

$$M = \frac{(n-1)n}{2} \quad \text{soit } 499\,500 \text{ pour l'exemple précédent}$$

Ces deux constatations permettent déjà de prévoir que les processus globaux tels que ceux résultant de certaines relations de LANCE et WILLIAMS (1967) ne peuvent pas être utilisés dans le cadre de cette étude car ils limiteraient de manière importante le nombre d'objets à classer ; alors que dans la réalité ce nombre n'est limité que par le temps. Dans ces processus la

limite serait en effet rapidement atteinte par un encombrement excessif de la mémoire rapide ou par un délai exagéré de la durée du traitement, si on utilise une mémoire de masse.

2) Analogie géométrique

Chaque objet peut être représenté par un point situé dans un espace à k dimensions qui correspondent aux variables descriptives. Il est alors possible de définir, entre deux objets, une relation d'analogie qui permet de les comparer.

L'ensemble formé par plusieurs objets peut être défini par le centre de gravité du groupe (barycentre) ; ce qui correspond à une notion de classe centroïde. Chaque groupe formé sera représenté par un point situé également dans l'espace à k dimensions.

La même relation d'analogie peut donc être utilisée :

- pour comparer deux points entre eux
- pour comparer un point à un groupe
- pour comparer deux groupes entre eux.

Cette analogie et les définitions qui en découlent permettent d'élaborer un processus séquentiel de traitement : en effet, si on condense l'information relative à chaque objet, au fur et à mesure du traitement, le nombre de positions mémoires dépend seulement du nombre de variables k et du nombre de groupes g : $M = k.g$.

Il est évident que le processus sera limité par le nombre de groupes, mais cette limitation est le cas de toutes les espèces biologiques.

3) Relations d'analogies

Employées couramment en taxonomie (SOKAL et SNEATH, 1963), (BOYCE, 1969), ces relations permettent d'effectuer les comparaisons envisagées précédemment, elles relèvent de deux catégories :

- Les relations de ressemblance (ou de similitude), auxquelles correspondent des coefficients qui vont de +1 à -1 ; le coefficient de corrélation ou le cosinus de l'angle de deux vecteurs appartiennent à cette catégorie.

- Les relations de proximité, auxquelles correspondent des fonctions variant de zéro - si l'identité est complète - à une valeur positive plus ou moins grande. Les fonctions de distance appartiennent à cette catégorie.

En raison de l'analogie géométrique et compte tenu du fait que certaines relations de ressemblance, tel que le coefficient de corrélation, entraînent une perte d'information dont il est difficile de présumer l'importance, nous avons utilisé des fonctions de distance.

- Métriques de MINKOWSKI

Les distances de MINKOWSKI (KRUSKAL, 1964) peuvent s'écrire selon la formule :

$$D_r(pq) = \left(\sum_{i=1}^k |x_{pi} - x_{qi}|^r \right)^{1/r}$$

$D_r(pq)$ étant la distance entre deux objets p et q

variables objets	1	i	k
1		⋮	
p	⋮	x_{pi}	⋮
q	⋮	x_{qi}	⋮
n			

si $r=1$, on a la somme des écarts absolus encore appelée "city-block distance" et qui sera notée D_1 . Lorsque les variables ne peuvent prendre que des valeurs égales à 0 ou à 1, on retrouve la distance de HAMMING.

si $r=2$, on a la distance euclidienne généralisée aux espaces à k dimensions, elle sera notée D_2 .

- Métrique hypersphérique

Pour résoudre certains problèmes concernant des relations homothétiques, nous avons utilisé également une fonction qui est égale à la longueur du grand arc qui sépare les points représentatifs des deux objets après projection sur une hypersphère de rayon 1. Cette métrique est un cas particulier des géométries de RIEMANN, deux objets p et q sont absolument identiques si leurs coordonnées sont reliées par la formule : $x_{pi} = x_{qi}$ pour $i=1,k$. Lorsque cette métrique est utilisée, le centre de gravité du groupe est également calculé sur l'hypersphère selon les expressions développées ci-après, la distance correspondante sera appelée "distance hypersphérique" et notée D_h .

Si p et q sont deux groupes d'effectifs n_p et n_q , les coordonnées cartésiennes y_{pi} et y_{qi} de leurs points représentatifs sur l'hypersphère de rayon 1 sont données par les formules :

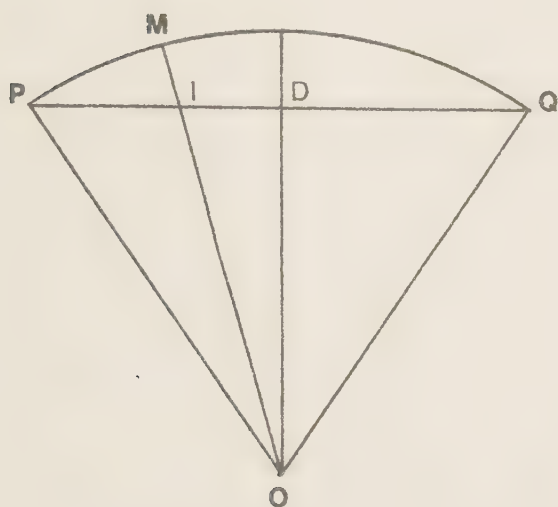
$$y_{pi} = \frac{x_{pi}}{\sqrt{\sum_{i=1}^k x_{pi}^2}} \quad \text{et} \quad y_{qi} = \frac{x_{qi}}{\sqrt{\sum_{i=1}^k x_{qi}^2}} \quad \text{pour } i = 1, k$$

la distance égale à la longueur de l'arc sur l'hypersphère est :

$$D_{h(pq)} = \text{arc cos} \left[\sum (y_{pi} \cdot y_{qi}) \right]$$

Pour calculer les coordonnées cartésiennes du point M représentant le centre de gravité sur l'hypersphère, il faut diviser l'arc PQ en deux arcs PM et MQ inversement proportionnels aux effectifs des groupes p et q $\widehat{POM} = \text{arc}(PM) =$

$$\text{arc}(PQ) \frac{n_q}{n_p + n_q}$$



puis on détermine la position du point I sur le segment PQ

$$\overline{OD} = \cos(\widehat{POQ}/2)$$

$$\overline{DI} = \overline{OD} \operatorname{tg}(\widehat{DOI}) = \overline{OD} \operatorname{tg}(\widehat{POM} - \widehat{POQ}/2)$$

$$\overline{PI} = \overline{PQ}/2 + \overline{DI}$$

$$\overline{IQ} = \overline{PQ} - \overline{PI}$$

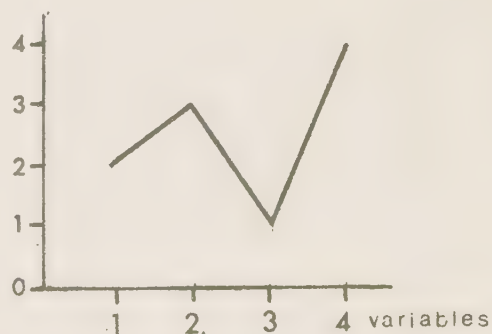
On calcule ensuite les coordonnées cartésiennes du point I à partir des coordonnées correspondant aux points P et Q en pondérant par PI et IQ, puis on calcule les coordonnées du point M en divisant chacune des coordonnées de I par le module $\overline{OI}$.

Transformations préalables des variables

Diverses transformations peuvent être effectuées avant le calcul de la distance, elles concernent : soit le profil de chaque objet (transformation en ligne), soit les variables (transformation en colonne).

a) Modification du profil

Le profil d'un objet est la figure obtenue en portant en abscisse les variables et en ordonnée la valeur prise par chaque variable, comme le montre le schéma ci-contre, pour un objet dont les valeurs respectives des quatre variables seraient : 2 3 1 4.



Selon DOZOL (1957), deux profils peuvent varier par la taille, la variabilité, ou la forme. Les différences de taille peuvent être éliminées en retranchant à chaque coordonnée la moyenne générale du profil.

$$x'_{pi} = x_{pi} - \bar{x}_p \quad \text{pour } i = 1, k \quad \text{et} \quad \bar{x}_p = \frac{\sum_{pi}^i x_{pi}}{k}$$

Les différences de variabilité s'éliminent en divisant chaque coordonnée par l'écart-type calculé sur l'ensemble des variables descriptives du profil :

$$x''_{pi} = x_{pi} / s_p \quad \text{pour } i = 1, k \quad \text{et} \quad s_p = \left[\frac{\sum_{pi}^i (x_{pi} - \bar{x}_p)^2}{k-1} \right]^{1/2}$$

Lorsque les différences de taille et de variabilité ont été éliminées, il reste les différences de forme :

$$x'''_{pi} = (x_{pi} - \bar{x}_p) / s_p \quad \text{pour } i = 1, k$$

Selon les transformations effectuées, deux objets p et q seront identiques si leurs coordonnées sont reliées par les formules :

transformation	formule
1 x'	$x_{pi} = x_{qi} + b$
2 x''	$x_{pi} = ax_{qi}$
3 x'''	$x_{pi} = ax_{qi} + b$

} \quad \text{pour } i = 1, k

L'utilisation de ces transformations est fonction du problème étudié et sera envisagée à propos des essais ou des applications. A l'examen des formules, il apparaît que la deuxième transformation est du même type que celle qui est effectuée pour la métrique hypersphérique ; il faut toutefois noter que le centre de gravité n'est pas calculé de la même façon.

b) Transformation sur les variables

Pour la plupart des variables biologiques, il est nécessaire de tenir compte du facteur d'échelle : sinon le simple fait de changer d'unité de mesure (en prenant un multiple ou un sous-multiple par exemple) modifierait profondément les résultats et la classification s'effectuerait alors uniquement sur les variables qui auraient la plus grande variabilité. Dans le cas des variables quantitatives, ce facteur peut être éliminé en divisant chaque variable par son écart-type calculé sur l'ensemble de l'échantillon.

$$x_{pi} = x_{pi} / s_i \quad \text{pour } i = 1, k \quad \text{et} \quad p = 1, n \quad s_i = \left[\frac{\sum_{pi}^p (x_{pi} - \bar{x}_i)^2}{n-1} \right]^{1/2}$$

DOZOL a démontré que la distance euclidienne calculée sur la matrice des notes en facteurs (obtenues après une analyse factorielle en composantes principales), pondérées par la racine carrée de la racine caractéristique des

facteurs correspondants, est la même que celle qui est obtenue à partir de la matrice des variables brutes préalablement standardisées.

Cette transformation fournit donc une méthode relativement simple d'application pour supprimer la redondance et attribuer un poids égal à chaque variable (conditions énoncées par ADAMSON dès 1757).

Il est également possible de pondérer différemment chaque variable, mais il est souvent difficile de trouver des critères pour attribuer les poids.

Cas des caractères qualitatifs - Echelles nominales

Deux types de codage peuvent être utilisés selon que l'échelle nominale correspond au caractère à deux niveaux ou plus :

- échelle nominale à deux niveaux. C'est le cas des caractères du type "tout ou rien" ; ils peuvent être codés par 0 ou 1 et les fonctions de distance de MINKOWSKI peuvent être utilisées. Dans ce cas, les éléments sont répartis sur un treillis de BOOLE et D_1 mesure le nombre de segments qu'il faut parcourir pour aller d'un élément à un autre.

- échelle nominale à plus de deux niveaux. Il est nécessaire de décomposer le caractère en autant de variables qu'il y a de niveaux.

Exemple : pour une couleur qui peut prendre les nuances, bleu, rouge ou jaune, on codera comme indiqué ci-dessous :

variable couleur	1	2	3
bleu	1	0	0
rouge	0	1	0
jaune	0	0	1

Dans les deux cas, la valeur correspondante du barycentre indique le pourcentage de présence du caractère au niveau considéré. Pour le calcul de la distance les deux types de codage ne sont pas équivalents, si on utilise D_1 une différence compte pour deux unités avec le second type et pour une avec le premier, il faudra donc pondérer en conséquence.

En fixant des coefficients de pondération adéquats, il est également possible de mélanger des variables qualitatives et quantitatives (PARK, 1968).

B. ELABORATION DU PROCESSUS

Compte tenu des buts fixés, la mise en oeuvre d'un programme de traite-

ment suppose que soient définis un processus de classification, d'identification et de structuration des groupes obtenus. Les choix effectués pour la mise au point de ces processus ont été fait essentiellement en fonction de la simplicité, du minimum d'encombrement mémoire et de la rapidité des calculs.

1) Processus de classification

Le "système identifiant" procède d'abord à la formation des classes à partir des premiers objets qui lui sont présentés. Pour cela, ses caractéristiques sont préalablement fixées par divers paramètres tels que, le nombre des variables "NV" qui servent à décrire les objets, le nombre de groupes "NG" à obtenir, ainsi que la fonction de proximité à utiliser. La taille de la mémoire mise ainsi à sa disposition est égale à $NV \cdot NG$.

Au cours d'une première phase, il prend les informations relatives aux NG premiers objets et les range dans les mémoires qui correspondent aux définitions des classes demandées. En fait, il forme donc NG classes à un objet.

Au cours de la deuxième phase, il classe successivement les nouveaux objets qu'on lui présente : pour cela il range, dans une mémoire de travail, l'information concernant le nouvel arrivant "O" ; puis il calcule la distance de cet objet au groupe le plus proche "G", ainsi que la distance minimale inter-groupe, soit "A" et "B" les groupes les plus voisins et D_{AB} leur distance.

- si $D_{OG} \leq D_{AB}$ le nouvel objet O est assigné au groupe G
- si $D_{OG} > D_{AB}$ les groupes A et B sont fusionnés et l'objet O forme l'embryon d'un nouveau groupe.

La figure 1 montre le schéma de principe du processus qui est également commenté à propos de l'essai n°1.

La recherche du groupe le plus proche nécessite le calcul de la fonction de proximité entre l'objet O et les NG groupes. La recherche de la distance minimale inter-groupe implique le calcul d'une matrice triangulaire de $NG(NG-1)/2$ éléments. Le temps nécessaire au classement d'un objet augmente donc avec le nombre de groupes demandés. En outre, l'assignation d'un objet nouveau à un groupe ou la fusion de deux groupes entraîne le calcul de nouvelles coordonnées barycentriques ainsi que de nouveaux effectifs.

Au fur et à mesure que se déroule le processus, les groupes grossissent

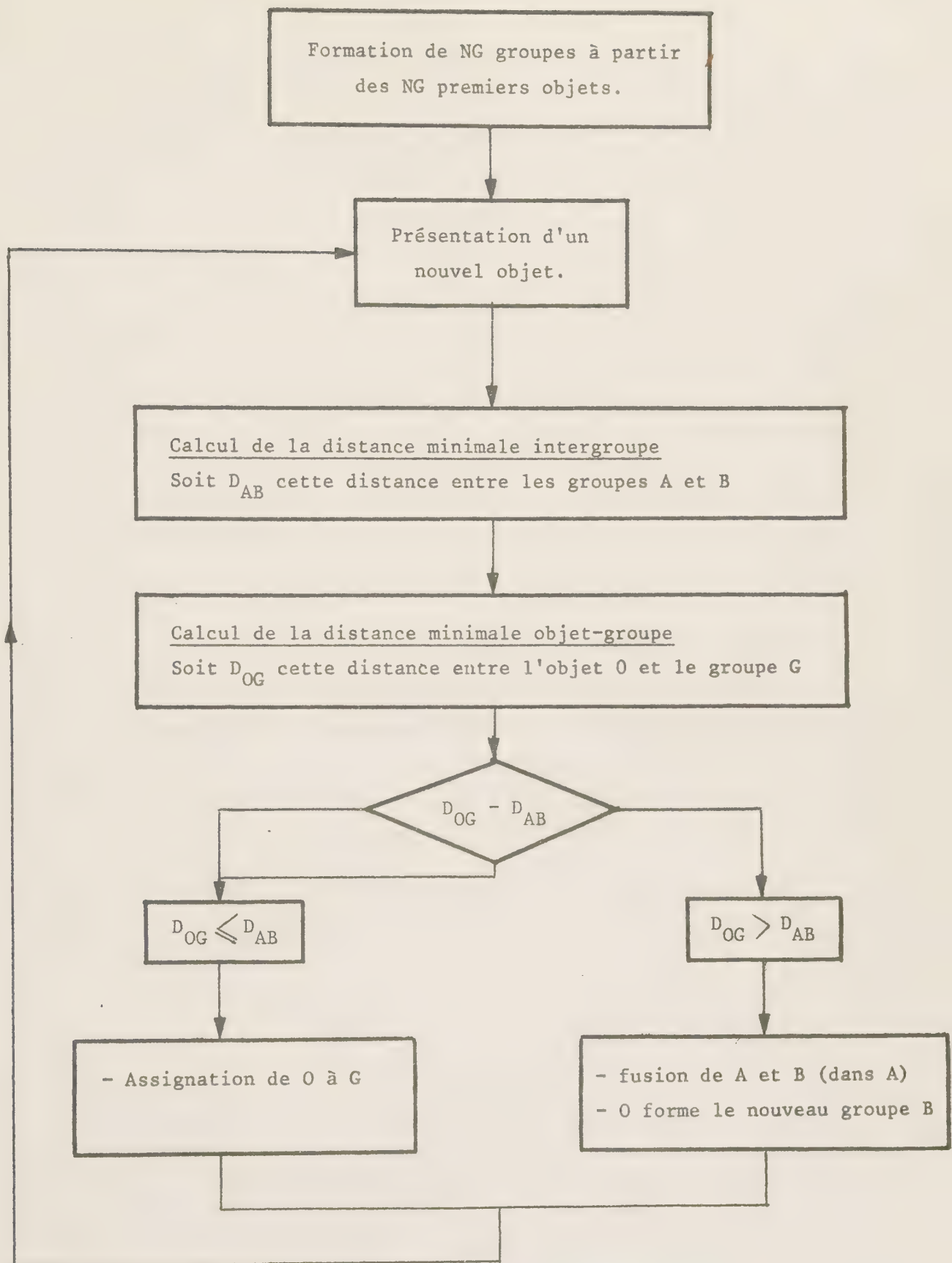


Figure 1 - Schéma de principe du processus de classification.

et les déplacements des centres de gravité, provoqués par l'adjonction de nouveaux objets, deviennent de plus en plus faibles. A partir d'un certain stade les groupes formés sont relativement stables ; on peut donc admettre que la formation des classes est terminée et passer au processus d'identification. L'arrêt du processus de classification est obtenu par un paramètre qui fixe le nombre d'objets nécessaires à la formation des classes. En fait, ce paramètre détermine la phase d'apprentissage du système : cela suppose qu'il ait, pendant cette période, traité au moins un objet de chacune des classes représentées dans l'échantillon. L'ordre de présentation des objets a aussi une influence sur le processus, si on estime que la classification est terminée lorsqu'il n'y a plus de fusion de classes mais uniquement des assignations d'objets nouveaux à des classes précédemment formées ; la convergence de la classification vers cette forme stable se produit d'autant plus vite que le système reçoit un assortiment d'objets appartenant à toutes les classes.

2) Processus d'identification

Le système classe ensuite les objets restant dans les groupes les plus proches, sans modifier les coordonnées barycentriques de ces groupes. La valeur de la fonction de proximité est indiquée, il est ainsi possible de détecter la présence d'un objet qui appartiendrait à une classe différente de celles qui ont servi à l'élaboration de la classification.

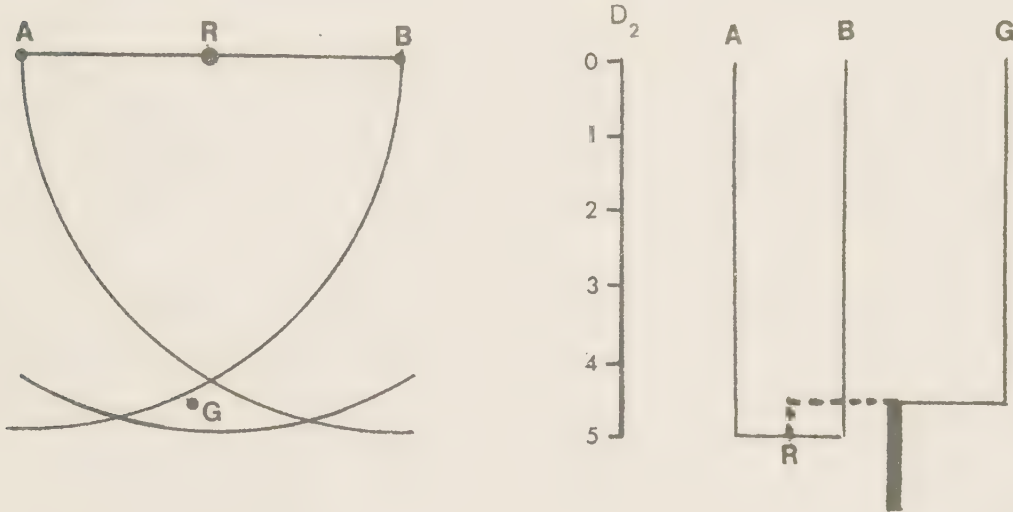
Le principal intérêt de ce processus est dû au gain de temps appréciable qui résulte du fait que la matrice de distance inter-groupe n'est plus calculée.

3) Processus de structuration - regroupement

Ce processus, mis en oeuvre à la fin du traitement, a pour but l'établissement d'un arbre de classification (appelé aussi dendrogramme), cet arbre permet de mettre en évidence la parenté qui existe entre les groupes obtenus. A chaque étape du regroupement, les deux groupes les plus proches sont fusionnés jusqu'à ce qu'il ne reste plus qu'un groupe unique qui correspond à l'ensemble de l'échantillon.

En général, la valeur de la fonction de proximité augmente à chaque fusion ce qui donne à l'arbre une structure hiérarchique. Mais, il peut arriver, comme le montre le schéma ci-joint, que la distance d'un groupe R, nouvellement formé à partir de deux groupes A et B, à un groupe G soit plus

courte que la distance qui existait entre A et B avant leur fusion. Dans ce cas, le dendrogramme présente une anomalie car la distance de fusion des groupes A et B (D_{AB}) est supérieure à la distance de fusion des groupes R et G (D_{RG}) bien que cette dernière se produise après.



Il existe diverses méthodes qui permettent d'éviter ces anomalies, certaines ne peuvent être employées ici en raison des choix effectués ; mais on peut utiliser la méthode de WARD qui effectue le regroupement sur le critère du minimum de déplacement (DUCIMETIERE, 1970). Le déplacement correspond à la somme des distances qui séparent les points contenus dans les groupes A et B du groupe R obtenu par la fusion de A et de B. Cette fonction, qui n'est pas une distance, peut être calculée par la formule :

$$\text{Déplacement}_{AB} = \frac{N_A \cdot N_B}{N_A + N_B} D_{AB}$$

N_A effectif du groupe A
 N_B effectif du groupe B
 D_{AB} distance euclidienne du groupe A au groupe B

Cette procédure a été introduite dans le programme ; elle permet d'éviter l'anomalie signalée précédemment, mais présente l'inconvénient de faire intervenir, de manière importante, les effectifs des groupes si bien que le regroupement peut, dans certains cas, ne plus tenir compte des ressemblances ce qui, sur le plan biologique, est un inconvénient majeur. De plus, ces anomalies se produisent dans des situations où les groupes sont très voisins et dans ce cas l'ordre de la fusion a peu de signification ; il est alors plus objectif de considérer que les trois groupes se sont fusionnés en même temps, plutôt que de chercher des différences plus ou moins arbitraires.

C. PRESENTATION DU PROGRAMME

1) Présentation des données

Les options et les paramètres nécessaires à la mise en oeuvre du programme de classification et d'identification "CLASSID" sont fournies par huit cartes-paramètres.

Après les cartes-paramètres, sont placées les cartes qui donnent les coefficients de division (écarts-types le cas échéant) et celles qui contiennent les coefficients de multiplication (pondération) si ces opérations ont été demandées.

Ensuite, sont placées les cartes qui correspondent à chaque objet. Ceux-ci doivent être identifiés par un numéro ne dépassant pas quatre chiffres et décrits par NV variables.

Le paquet de cartes données comprend donc :

- les cartes paramètres
- les cartes du diviseur
- les cartes du multiplicateur
- les cartes de description des objets

Cartes paramètres

carte 1 commentaires

- numéro de la carte col. 3
- commentaires col. 4 à 79

Carte 2 format diviseur

- numéro de la carte col. 3
- format de lecture du diviseur (1) col. 4 à 79

Carte 3 format multiplicateur

- numéro de la carte col. 3
- format de lecture du multiplicateur (1) col. 4 à 79

Carte 4 format des NG premiers objets

- numéro de la carte col. 3
- format de lecture (2) col. 4 à 79

Carte 5 format de lecture des objets suivants

- numéro de la carte col. 3
- format de lecture (2) col. 4 à 79

Carte 6 nombre d'objets et nombre de groupes

- numéro de la carte col. 3
- nombre de groupes demandés NG col. 7 et 8
- limite du processus de classification LC col. 9 à 13
- nombre d'objets à traiter NO col. 14 à 18

Carte 7 caractéristiques du traitement

- numéro de la carte col. 3
- nombre de variables NV col. 7 et 8
- formule de la relation de proximité col. 13
 - 1 somme des écarts absolus..... D_1
 - 2 distance euclidienne D_2
 - 3 distance hypersphérique D_h
- demande de division préalable col. 18
- demande de multiplication préalable col. 23
 - 0 l'opération n'est pas demandée
 - 1 l'opération est demandée

Carte 8 caractéristiques de classification et d'impression

- numéro de la carte col. 3
- critère de classification col. 8

	processus	classification	regroupement
0		distance	distance
1		distance	déplacement
2		déplacement	déplacement
- demande d'impression des matrices de distances inter-
groupes lors du regroupement col. 13
 - 0 pas d'impression
 - 1 impression

Ecriture des formats de lecture

- (1) lecture du diviseur ou du multiplicateur
 - faire lire NV nombres réels en mode F ou E
- (2) lecture des objets
 - faire lire 1 nombre entier en mode I pour le
numéro d'identification
 - puis faire lire NV nombres réels en mode F
ou E pour les NV variables descriptives.

Dimensions maximales

Les dimensions maximales dépendent des caractéristiques de l'ordinateur employé, les dimensions indiquées ci-après sont celles qui ont été utilisées pour l'ensemble des travaux exposés ici.

- taille de la mémoire de stockage des informations concernant la description des groupes $M = NV \cdot NG \leq 1200$
- nombre de groupes demandés $NG \leq 30$
- nombre de variables $NV \leq 70$
- nombre d'objets à traiter $NO \leq 200$

Le programme est conçu de telle manière que la mémoire M est automatiquement répartie (en fonction des paramètres) en secteurs correspondant à chaque groupe et cela quel que soit le nombre des groupes et des variables. Avec une mémoire M égale à 2000, il serait possible de rechercher 20 groupes sur des objets définis par 100 variables ou 10 groupes si le nombre des variables est 200. La taille de cette mémoire est donc le principal facteur qui limite les dimensions des problèmes à traiter : les autres facteurs intervenant relativement peu, leurs limites peuvent être facilement augmentées.

2) Présentation des résultats

Le programme imprime les cartes-paramètres puis, pendant les processus de classification et d'identification, il indique au fur et à mesure les assignations et fusions avec les distances correspondantes.

Lorsque tous les objets ont été classés, il donne :

- la composition des groupes : pour chaque groupe sont indiqués les identificateurs des objets classés dans le groupe
- la description des groupes, c'est-à-dire les coordonnées du barycentre
- la matrice de distance inter-groupe.

Lors du regroupement, à chaque étape, sont indiquées :

- la fusion effectuée : le nouveau groupe prend le plus petit numéro des groupes dont il provient

- la description du nouveau groupe obtenu
- la matrice de distance inter-groupe si elle a été demandée par les cartes paramètres

Remarques

1) Pour les métriques de MINKOWSKI (distance D_1 et D_2), les opérations préalables (division ou multiplication) sont effectuées lors du calcul des distances ; il en résulte que les coordonnées correspondant à la description des classes sont indiquées dans les échelles d'origine ce qui facilite l'exploitation des résultats.

2) Pour la métrique hypersphérique (distance D_h), les descriptions correspondent à un profil commun aux objets homothétiques. Le programme donne le profil de chaque objet et ceux des groupes obtenus.

D. ESSAIS EFFECTUES

Trois essais ont été effectués pour la mise au point du programme ; ils ont également permis de formuler, de manière simple, divers types de problèmes et de tester le processus dans des conditions correspondant aux hypothèses fondamentales.

ESSAI N° 1 - Utilisation de la distance euclidienne D_2

Un échantillon de 35 points a été formé en répartissant des points dans un espace à deux dimensions, de manière à obtenir sept groupes, comme l'indique le schéma A de la figure 2. Plusieurs traitements ont été effectués sur cet échantillon, et les tableaux qui suivent donnent le résultat d'un essai pour lequel sept groupes avaient été demandés. Le schéma 2B donne l'arbre de classification correspondant, le schéma 2C montre une série de sept objets qui pourraient correspondre aux points considérés. Sur la figure 3, l'évolution de quelques distances caractéristiques a été représentée ; on remarquera que les distances se situent dans des zones différentes selon qu'elles correspondent à des assignations, à des fusions ou à des regroupements.

Pour tester l'influence du nombre de groupes demandés, un traitement a été effectué en demandant douze groupes. La figure 4 illustre les résultats obtenus, les schémas 4B et 4C montrent qu'au cours du regroupement les sept groupes sont reconstitués et qu'il existe une zone ($D_2 \leq 4$) où les sous-groupes obtenus se rassemblent, la partie inférieure du dendrogramme n'étant pas modifiée.

PROGRAMME C L A S S I D

1 ESSAI N° 1 - ECHANTILLON DE 35 POINTS DECRITS PAR 2 VARIABLES
2
3
4(1X,I2,2F5.1)
5(1X,I2,2F5.1)
6 7 35 35
7 2 2 0 0
8 0 1

FORMATION DES GROUPES

FUSION DE DEUX GROUPES

ASSIGNATION A UN GROUPE

ECART	GRUPE	GRUPE	EFF.	*	ECART	INDIVIDU	GRUPE	EFF.
				*		22 ->	1	1
				*		31 ->	2	1
				*		21 ->	3	1
				*		35 ->	4	1
				*		9 ->	5	1
				*		10 ->	6	1
				*		11 ->	7	1
1.414	3 ->	1	2	*	1.118	13 ->	6	2
1.953	7 ->	6	3	*		14 ->	3	1
				*	1.581	15 ->	7	1
2.088	6 ->	3	4	*		16 ->	7	2
				*	1.275	17 ->	6	1
				*	1.581	18 ->	7	3
2.173	7 ->	6	4	*		7 ->	5	2
				*	1.186	1 ->	7	1
				*	.537	19 ->	6	5
				*	1.275	12 ->	3	5
2.500	4 ->	2	2	*		8 ->	5	3
				*		29 ->	4	1
				*	1.414	30 ->	4	2
				*	2.000	2 ->	7	2
				*	1.414	3 ->	7	3
				*	1.700	4 ->	7	4
				*	1.864	5 ->	5	4
				*	1.305	6 ->	5	5
				*	2.550	23 ->	1	3
				*	.687	24 ->	1	4
				*	1.505	25 ->	1	5
				*	2.121	26 ->	4	3
				*	1.031	32 ->	2	3
				*	.687	33 ->	2	4
				*	1.581	27 ->	4	4
				*	.395	28 ->	4	5
				*	1.924	20 ->	6	6
				*	2.183	34 ->	2	5

IMPRESSION DES GROUPES

GROUPE		INDIVIDUS					
E	1	22	21	23	24	25	
G	2	31	35	32	33	34	
C	3	10	11	13	14	12	
F	4	29	30	26	27	28	
B	5	9	7	8	5	6	
D	6	15	16	17	18	19	20
A	7	1	2	3	4		

DESCRIPTION DES GROUPES

GROUPE		EFF.	x_1	x_2
E	1	5	6.800	5.600
G	2	5	16.100	2.200
C	3	5	14.220	19.800
F	4	5	12.100	7.300
B	5	5	6.300	21.000
D	6	6	16.417	15.250
A	7	4	3.750	17.250

DISTANCES INTERGROUPES

	E	G	C	F	B	D
	1	2	3	4	5	6
G 2	9.902					
C 3	16.022	17.700				
F 4	5.566	6.481	12.679			
B 5	15.408	21.201	8.010	14.877		
D 6	13.624	13.054	5.053	9.046	11.637	
A 7	12.043	19.469	10.776	12.989	4.535	12.824

REGROUPEMENT	CRITERE	1
--------------	---------	---

REGRUPPEMENT

NO	ELART	DANS	DESCRIPTION			
1	4.535	7 ->	5	9	5.167	19.333

DISTANCES INTERGROUPES

	1	2	3	4	5
2	9.902				
3	16.022	17.700			
4	5.566	6.481	12.679		
5	13.830	20.325	9.065	13.888	
6	13.624	13.054	5.053	9.046	11.968

REGROUPEMENT

NO	ELART	DANS			DESCRIPTION	
2	5.053	6 ->	3	11	15.418	17.318

DISTANCES INTERGROUPEs

	1	2	3	4
2	9.902			
3	14.546	15.134		
4	5.566	6.481	10.553	
5	13.830	20.325	10.448	13.888

REGROUPEMENT

NO	ECART	DANS	DESCRIPTION
3	5.566	4 -> 1 10	9.450 6.450

DISTANCES INTERGROUPES

	1	2	3
2	7.892		
3	12.399	15.134	
5	13.677	20.325	10.448

REGROUPEMENT

NO	ECART	DANS	DESCRIPTION
4	7.892	2 -> 1 15	11.667 5.033

DISTANCES INTERGROUPES

	1	3
3	12.845	
5	15.708	10.448

REGROUPEMENT

NO	ECART	DANS	DESCRIPTION
5	10.448	5 -> 3 20	10.805 18.225

DISTANCES INTERGROUPES

1

3 13.220

REGROUPEMENT

NO	ECART	DANS	DESCRIPTION
6	13.220	3 -> 1 35	11.174 12.571

ARBRE DE CLASSIFICATION

1	4.535	7 ->	5
2	5.053	6 ->	3
3	5.566	4 ->	1
4	7.892	2 ->	1
5	10.448	5 ->	3
6	13.220	3 ->	1

DENDROGRAMME

1	4	2	3	6	5	7
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*****	*
*	*	*	*	*	*	*
*	*	*	*	*****	*	*
*	*	*	*	*	*	*
*****	*	*	*	*	*	*
*	*	*	*	*	*	*
*****	*	*	*	*	*	*
*	*	*	*	*	*	*
*****	*	*	*	*	*	*
*	*	*	*	*	*	*

Figure 2 - Essai n° 1

Classification sur un échantillon de 35 points situés dans un espace à 2 dimensions.

Schéma A

Position des points et des groupes obtenus.

Les points sont repérés par des numéros variants de 1 à 35.
Les groupes sont repérés par des lettres, les numéros de groupe attribués par le programme sont arbitraires, ils ont été également indiqués.

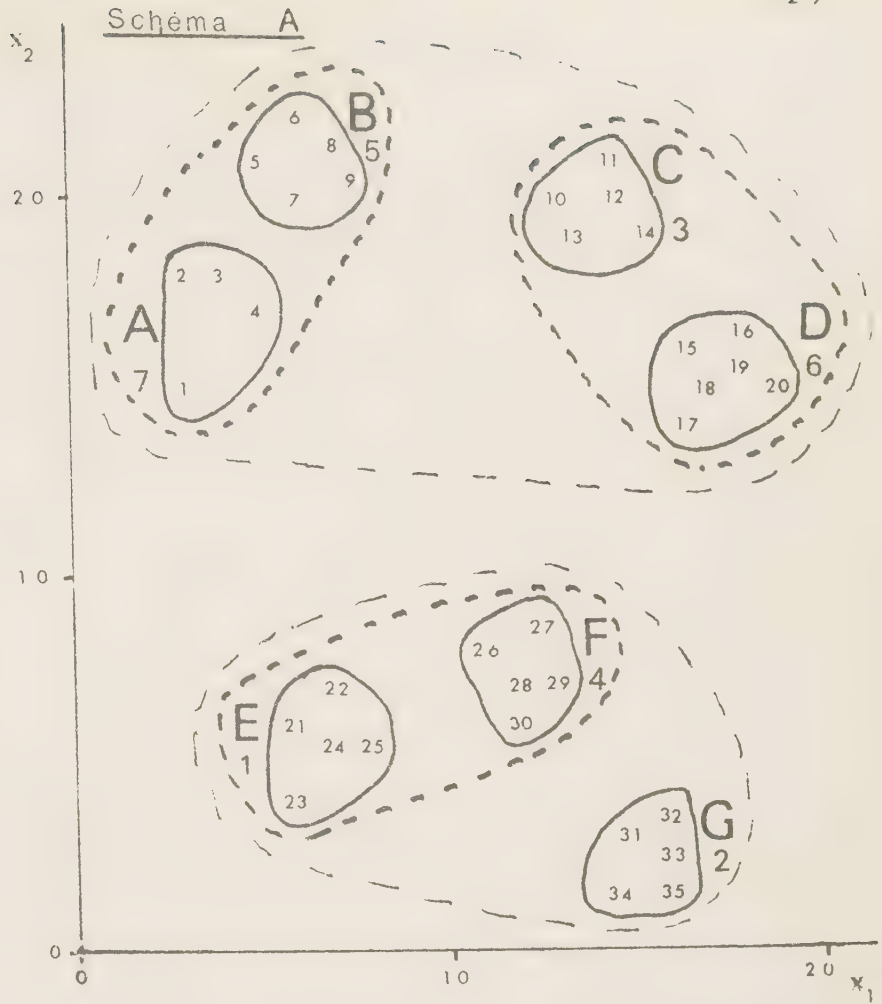


Schéma B

Arbre de classification (dendrogramme) indiquant les regroupements successifs en fonction de la distance.

Schéma B

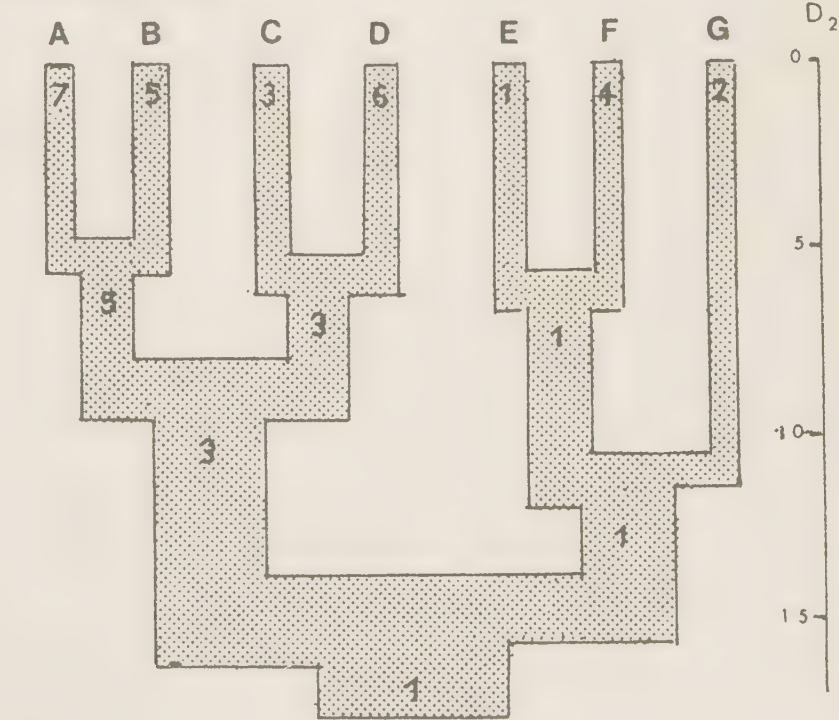
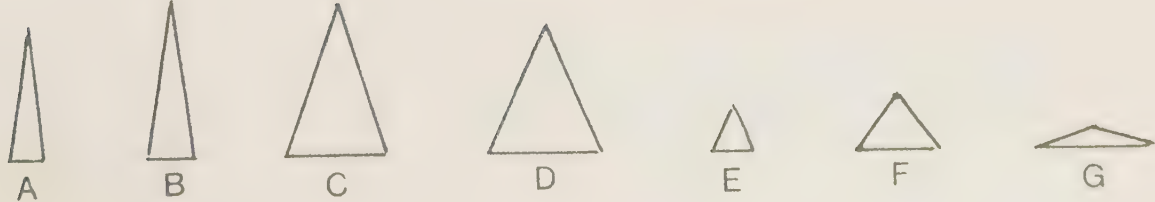


Schéma C

Si chaque point correspond à un triangle isocèle, x_1 étant la base et x_2 étant la hauteur, les triangles "moyens" de chaque groupe sont représentés sur le schéma.



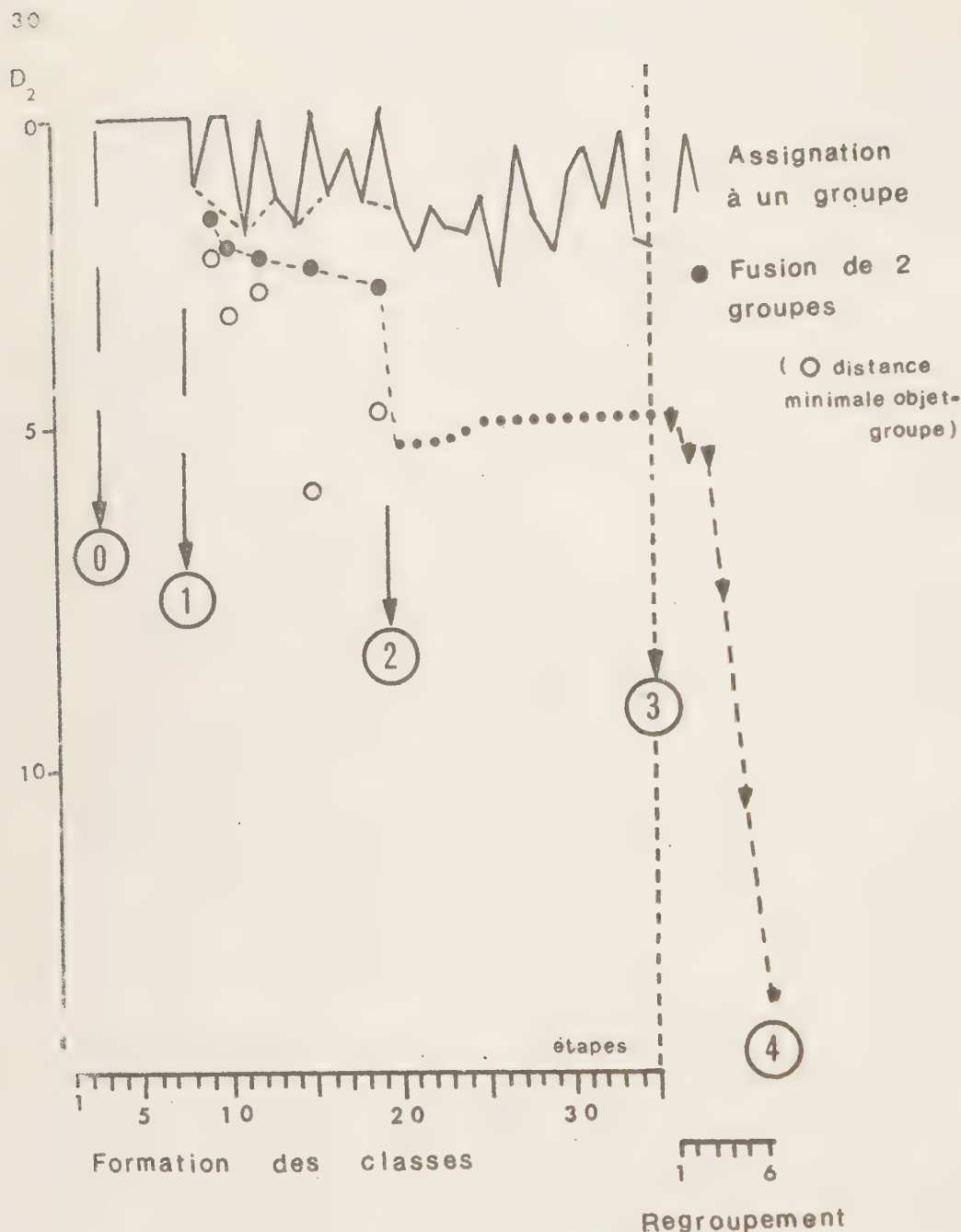


Figure 3 - Essai n° 1

Echantillons de 35 objets formant 7 Classes. Evolution de quelques distances caractéristiques au cours du processus de formation de classes et du regroupement.

Phase 0-1 : Prise d'information initiale. Les 7 premiers objets forment chacun une classe.

Phase 1-3 : Formation des classes. A chaque étape un nouvel objet est présenté et classé. Lorsque la distance minimale OBJET-GROUPE est inférieure à la distance minimale INTER-GROUPE l'objet est assigné au groupe correspondant; dans le cas contraire les deux groupes les plus proches sont fusionnés et le nouvel objet forme un nouveau groupe. A partir de la 19ème étape (repère 2) le programme a reçu un objet de chaque groupe, toutes les classes étant trouvées il n'y aura plus de fusion de groupes pendant la durée du processus de formation des classes.

Schéma A

Ce schéma montre les groupes obtenus. Les points noirs correspondent aux 12 points qui ont été donnés en premier.

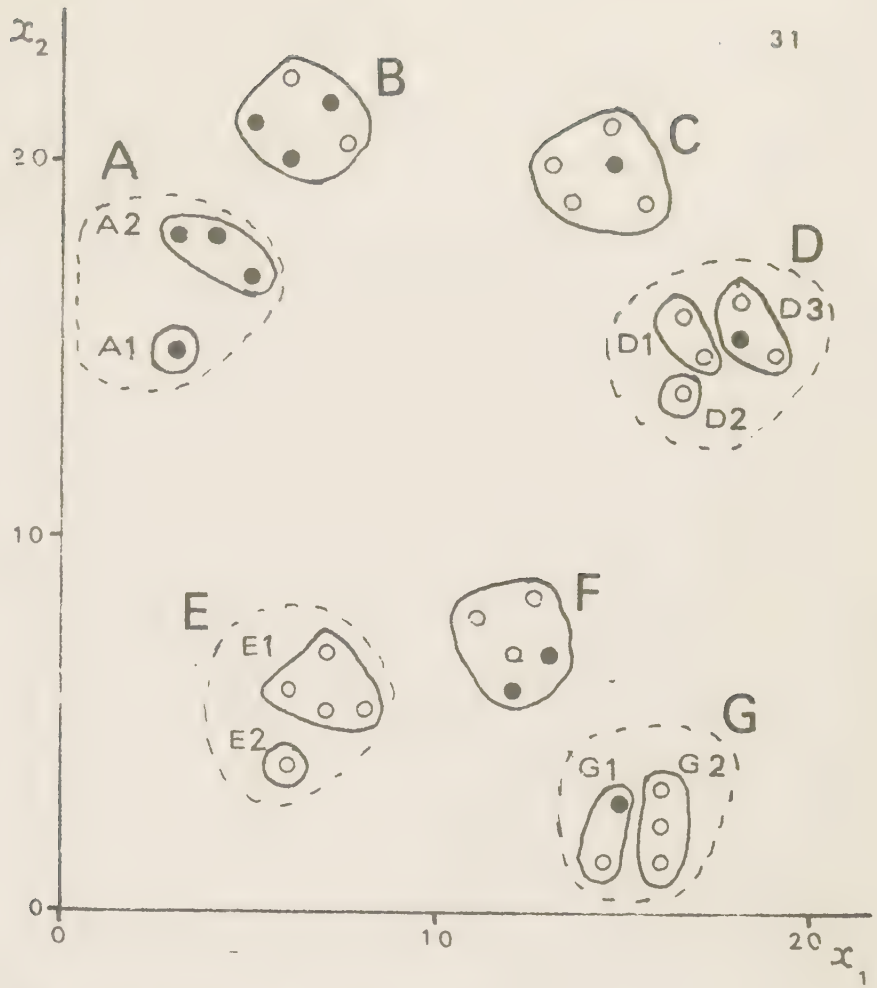


Schéma B

Arbre de classification (Dendrogramme).

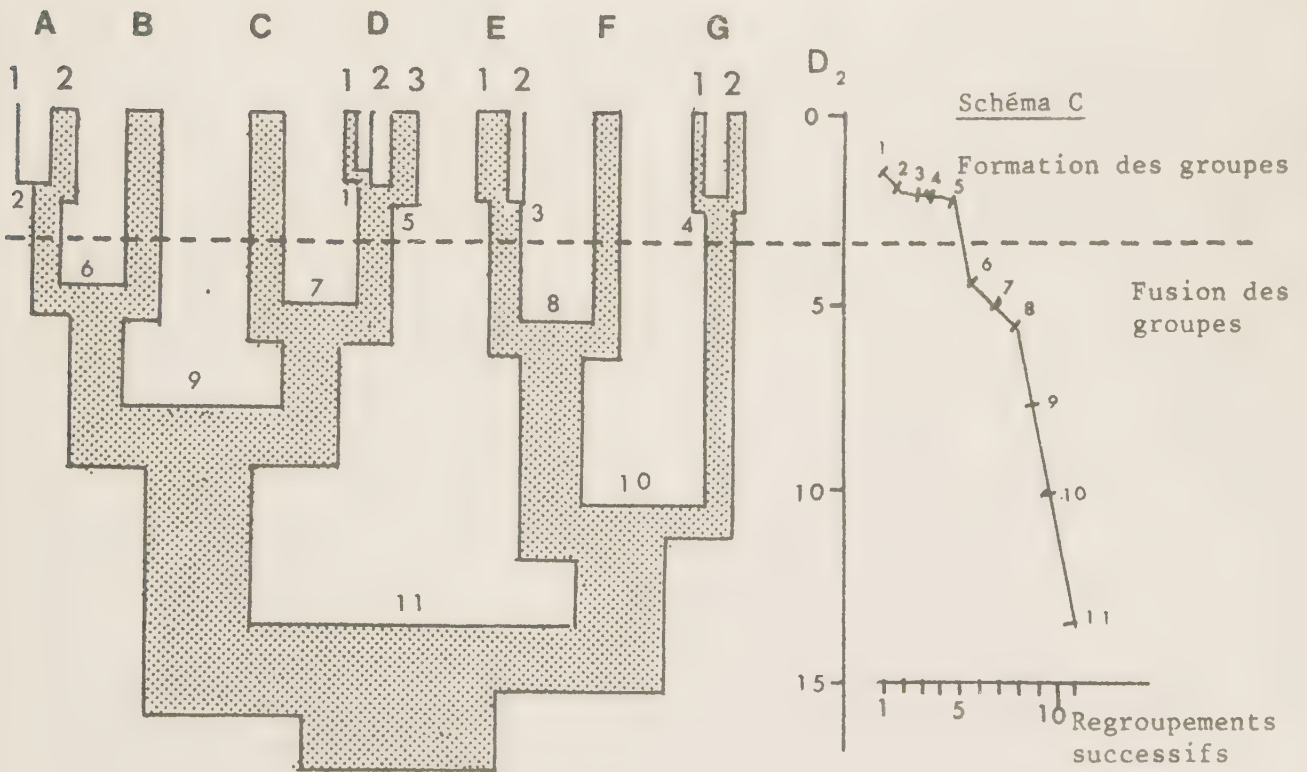


Schéma C

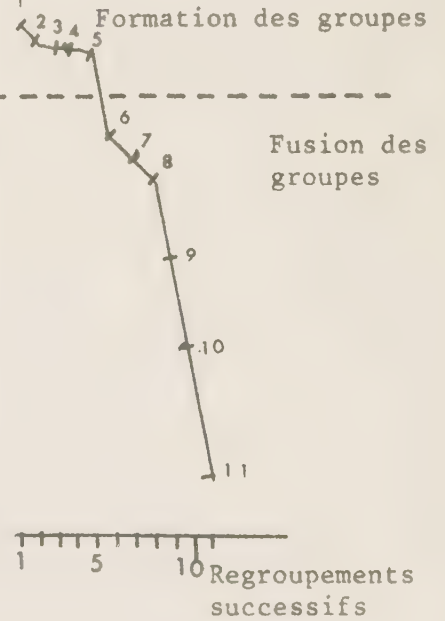


Figure 4 - Influence du nombre de groupes demandés.

Si on augmente le nombre de groupes la discrimination devient plus grande. Les schémas B et C font apparaître deux zones de regroupement qui correspondent à la formation des groupes et à leur fusion.

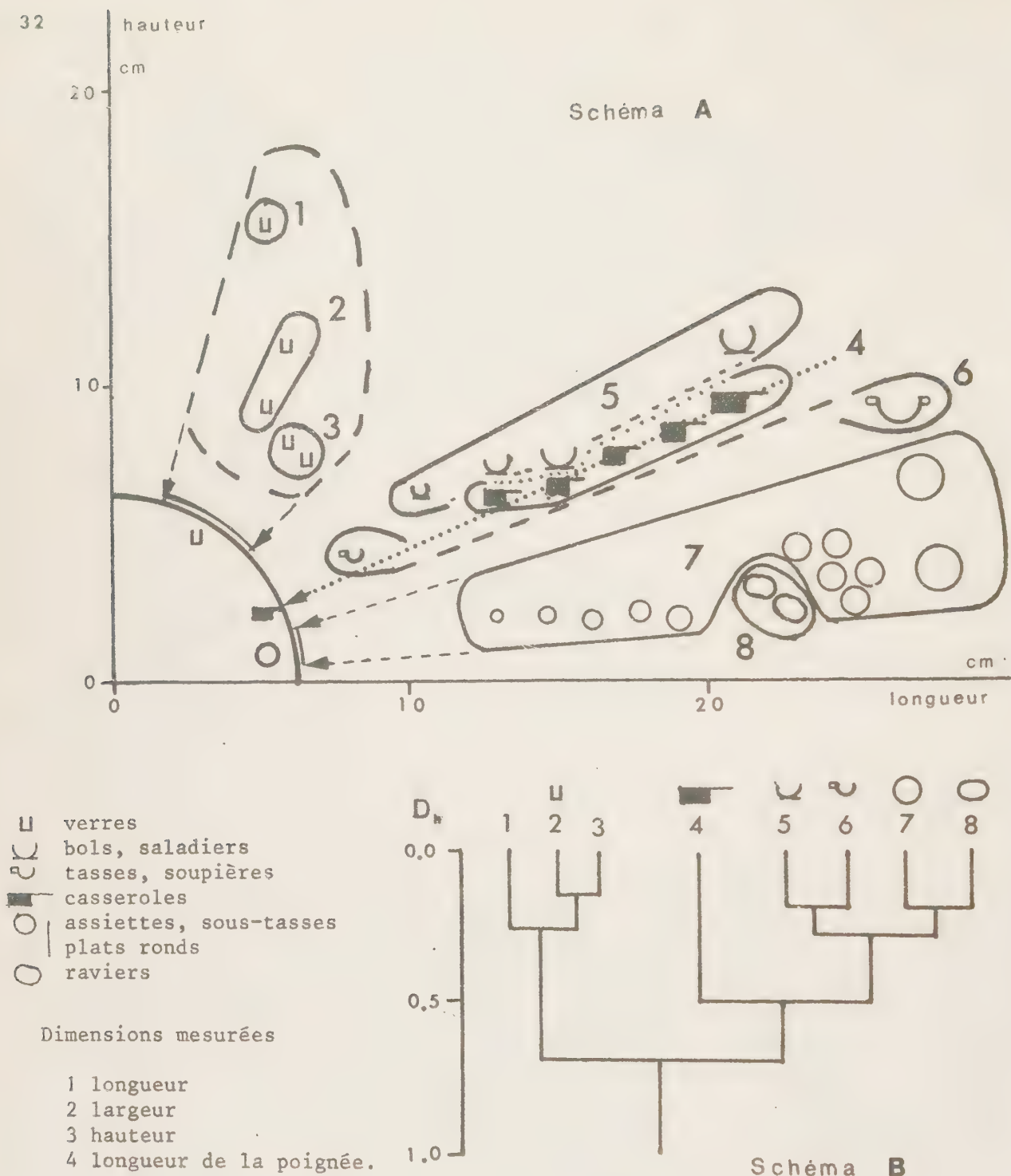


Figure 5 - Essai n° 2

Principe de la reconnaissance de forme de 30 objets usuels de dimensions homothétiques.

- A - Représentation à deux dimensions (longueur et hauteur) des 30 objets. Le processus de classification projette d'abord l'objet sur une hypersphère de rayon 1 (à deux dimensions c'est un cercle) et le classe ensuite.
- B - Classification obtenue à partir des quatre variables mesurées. Les huit groupes demandés correspondent aux branches du dendrogramme et son représentés sur le schéma A.

ESSAI N° 2 - Utilisation de la métrique hypersphérique (distance D_h)

Cet essai a été réalisé sur un échantillon de 30 objets usuels choisis parce qu'ils présentaient des relations homothétiques. Le schéma 5A montre la projection des objets sur un espace à deux dimensions et illustre la manière dont les groupes sont constitués. La présence des deux autres variables est indispensable pour obtenir une bonne discrimination comme le montre les classes qui ont été reconstituées sur le schéma.

ESSAI N° 3 - Utilisation des caractères qualitatifs (distance D_1)

Le matériel d'étude pour cet essai a été obtenu par une démarche similaire à celle utilisée par SOKAL (1966) ; il est ainsi possible d'étendre la notion de groupe centroïde à des objets décrits par des variables qualitatives.

Un échantillon de 35 individus décrits par 20 variables qualitatives du type "tout ou rien" a été constitué de telle manière que les individus soient répartis en 5 groupes de 7. Pour cela, on a d'abord formé 5 individus-types relativement éloignés les uns des autres (D_1 variant de 8 à 15), puis les groupes ont été constitués autour de ces individus-types de telle façon qu'un individu diffère au plus par une seule variable du centre du groupe.

L'ensemble de l'échantillon a été classé : il y a lieu de remarquer qu'il est possible de retrouver les groupes et de présumer de leurs caractéristiques même si l'individu-type ne figure pas dans l'échantillon.

La figure 6 montre quelques-uns des groupes réalisés ; elle a été obtenue en attribuant une signification aux variables descriptives.

Schema A

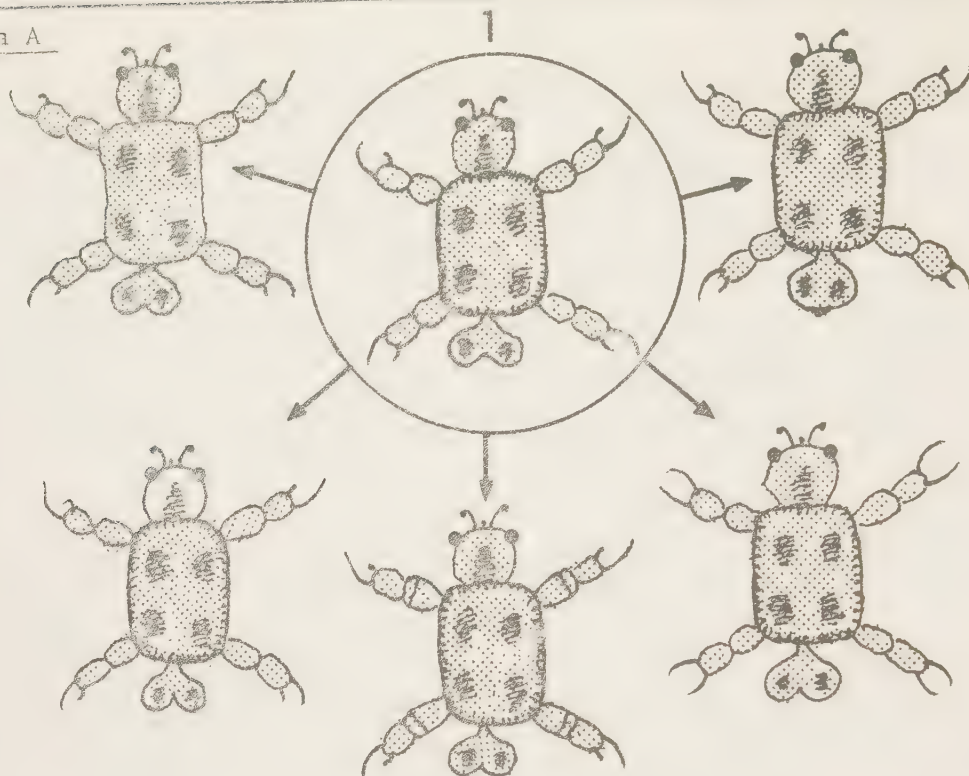


Schéma B

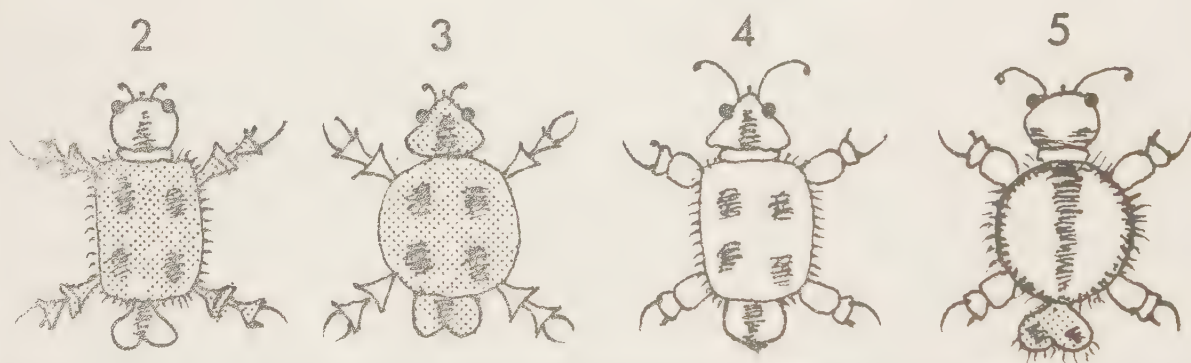


Schéma C

Groupe	1	2	3	4	5
1	0	8	8	13	11
2	8	0	10	9	13
3	8	10	0	11	15
4	13	9	11	0	10
5	11	13	15	10	0

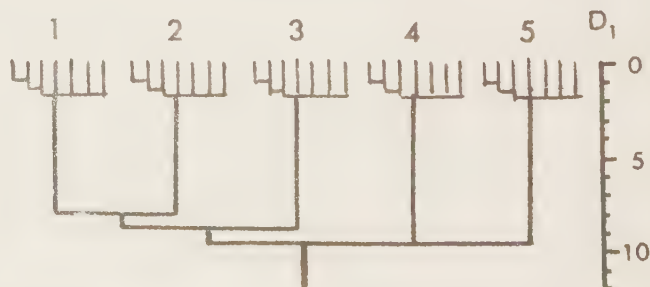


Figure 6 - Essai n° 3 - Utilisation des caractères qualitatifs.

Principe du codage et de la formation des classes sur un exemple de 35 individus appartenant à 5 groupes et décrits par 20 variables qualitatives.

Le schéma A montre 5 individus appartenant au groupe 1; ils diffèrent de l'élément-type (représenté dans le cercle) par un seul caractère. Les individus-types correspondant aux autres groupes sont représentés sur le schéma B; le schéma C donne les distances entre les individus-types et le dendrogramme obtenu lors de la classification.

III. APPLICATIONS

Diverses applications ont été entreprises pour essayer d'apporter une solution aux problèmes posés par les chercheurs. Elles concernent des études pour lesquelles une expérimentation a été effectuée et sur lesquelles différents traitements ont été faits à partir des données recueillies. Les applications exposées ici concernent le domaine de la microbiologie et celui de la biologie clinique.

A. APPLICATIONS RELEVANT DE LA MICROBIOLOGIE

Ces travaux ont été entrepris en vue de normaliser les épreuves d'identification des bactéries ou des levures et de simplifier leur mise en oeuvre à l'aide de microméthodes.

A plus long terme, le traitement automatique des informations recueillies par ces microméthodes est envisagé et les travaux exposés se situent dans cette perspective (JACQ, DEFAYOLLE et BUISSIÈRE, 1971), (DEFAYOLLE, COLOBERT..., 1968). Il est notamment apparu que l'identification des bactéries, à partir des méthodes nouvelles, ne pouvaient pas être efficacement envisagée sans aborder en même temps le problème de la formation des classes.

Les trois études réalisées ont porté sur les bactéries du genre *Lactobacillus*, de la tribu des *Klebsiellae* et sur un échantillon de levures. Pour les *Lactobacilles* et pour les *Klebsiellae*, nous avons utilisé les résultats obtenus respectivement par PAULE (1971) et par BARBE (1969) avec les microméthodes étudiées par BUISSIÈRE et coll. (1968 et 1969). Les aspects microbiologiques de ces travaux ayant fait l'objet de communications de leurs auteurs, seuls seront rappelés ici les aspects méthodologiques. Les travaux sur les levures ont été effectués avec CHARPENTIER qui a choisi l'échantillon d'étude ainsi que les variables.

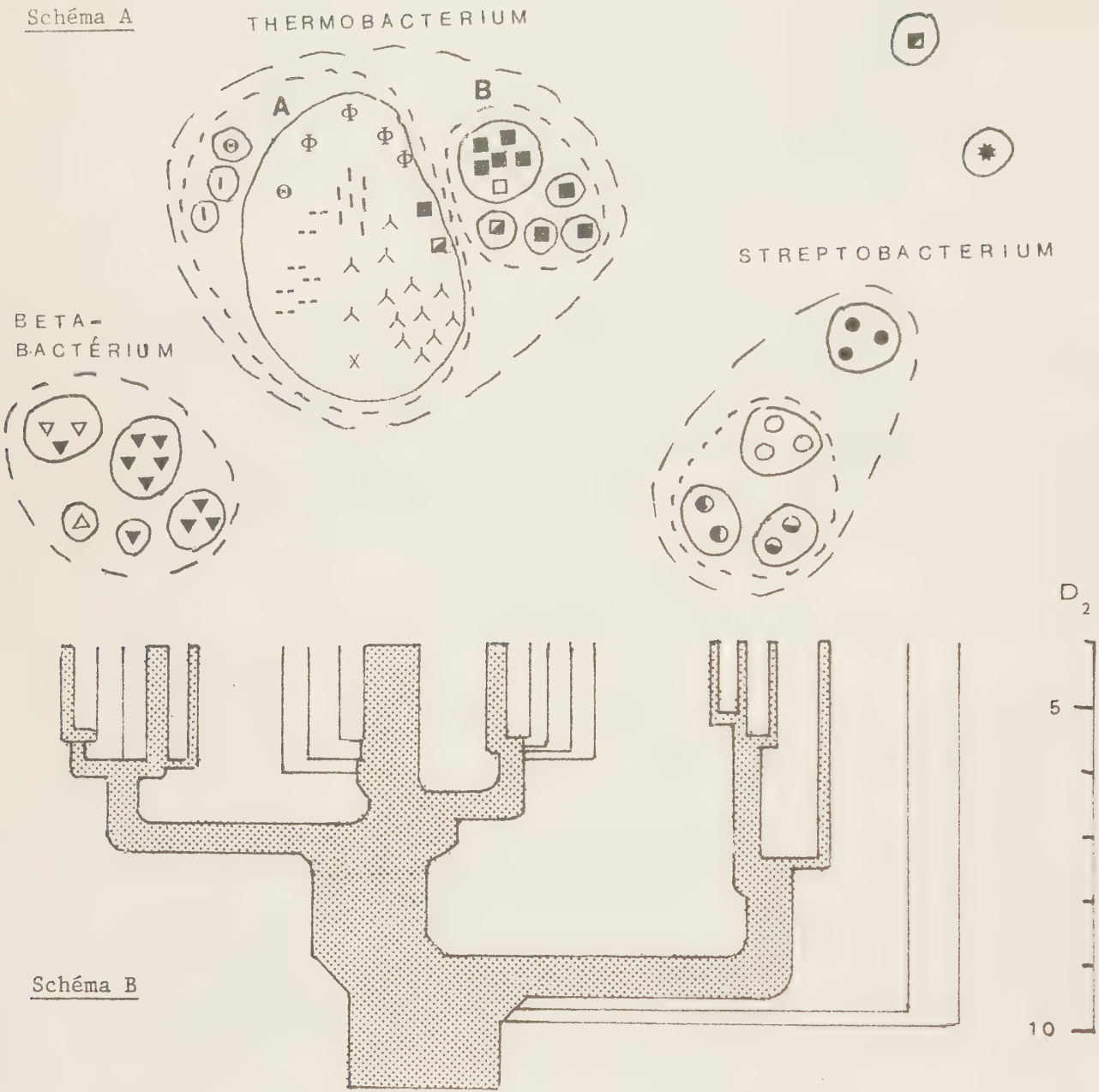


Figure 7 - Classification de 74 biotypes appartenant à 18 espèces du genre LACTOBACILLUS.

BETABACTERIUM

▼ fermenti...	10
▽ buchneri ..	2
△ brevis	1

THERMOBACTERIUM

	A	B
Λ lactis.....	13	
! bulgaricus .	9	
-- helveticus .	7	
Φ jugurti	4	
⊕ delbrückii .	2	
X viridescens.	1	
■ salivarius		
salivarius.	1	1
■ acidophilus.	1	8
□ leichmanii .		1

STREPTOBACTERIUM

casei	
⊙ alactosus...	2
⊙ casei	2
○ rhamnosis ..	3
● plantarum ..	3
■ salivarius	
salicinus .	1
* cellobiosus.	1

Description sommaire de la microméthode

Pour les germes étudiés ici, les auteurs ont mis au point une série de réactions biochimiques. Après culture sur un milieu adéquat, un certain volume du milieu de culture est déposé dans chacun des godets d'une plaque ; chaque godet correspondant à un substrat différent. Chaque plaque, placée dans une étuve, est ensuite examinée à 3, 6, 12 et 24 heures et les résultats des réactions biochimiques, qui se traduisent par des modifications de coloration, sont codés de 1 à 5 suivant la rapidité de la réaction.

1) Travaux effectués sur les Lactobacilles

Ces travaux ont porté sur un échantillon de 74 biotypes appartenant à 18 espèces du genre *Lactobacillus*, 47 variables ont été retenues.

La figure 7 représente les résultats obtenus, après traitement, par le programme Classid (20 groupes demandés, distance D_2). Sur le schéma 7A, nous avons essayé, en utilisant les distances calculées par le programme, de donner une idée de la composition et de la position relative des groupes formés. Cette représentation ne peut être que très approximative : les groupes se situant dans un espace à plus de deux dimensions. Le dendrogramme fait nettement apparaître les trois sous-genres *Betabacterium*, *Thermobacterium* et *Streptobacterium*. Les groupes obtenus dans les *Streptobacterium* correspondent exactement aux espèces définies par les microbiologistes ; les *Betabacterium* comprennent surtout des *L. fermenti* ; tandis que les *Thermobacterium* se répartissent en deux rameaux : A avec les espèces *L. lactis*, *bulgaricus*, *helveticus*, *jugurti* et B avec les *L. acidophilus*. Dans l'ensemble, ces résultats recoupent ceux qui ont été obtenus par les méthodes usuelles et qui sont exposés dans la thèse de PAULE (1971).

2) Travaux effectués sur les Klebsielleae

Ces études ont été réalisées sur un échantillon de 137 biotypes appartenant à 11 espèces de la tribu des Klebsielleae, 49 variables ont été utilisées.

La figure 8 représente les résultats obtenus ; le schéma 8A (extrait de la thèse de BARBE parue avant ce travail, 1969), représente la disposition relative des espèces dans l'espace des deux premiers facteurs. Le schéma 8B montre l'arbre de classification construit, après traitement, par le programme Classid. La concordance entre ces deux séries de résultats, obtenus par des méthodes différentes, est donc très satisfaisante. Cela n'était pas

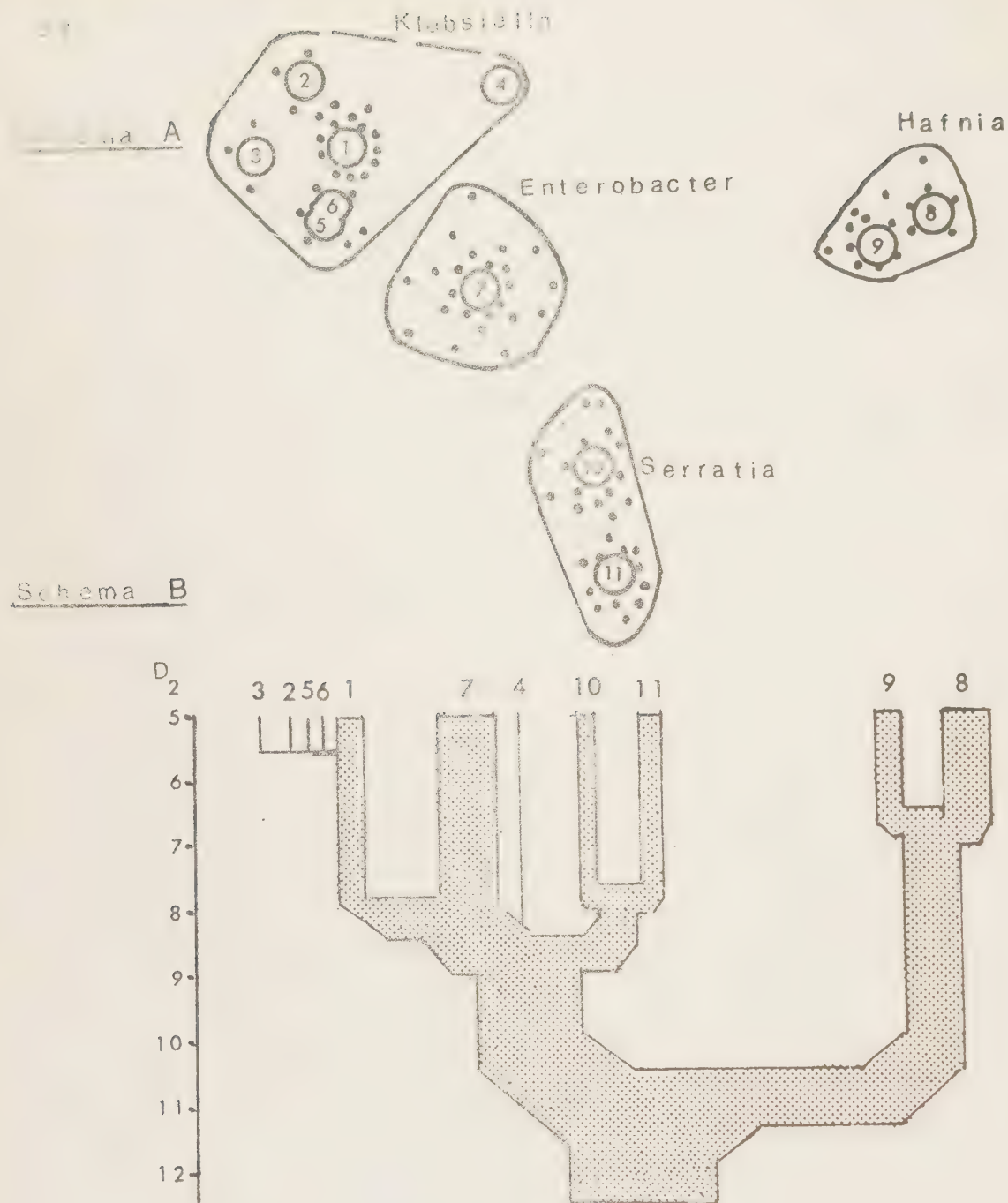


Figure 8 - Classification de 137 biotypes de la tribu des Klebsiellae.

A - Représentation des biotypes dans l'espace des deux premiers facteurs.

B - Classification obtenue à partir de 49 variables enzymatiques.

GENRE	CODE	ESPECE	NOMBRE
Klebsiella	1	Klebsiella pneumoniae	15
	2	Klebsiella oxytoca	1
	3	Klebsiella ozanae	1
	4	Klebsiella rhinoscleromatis	1
Enterobacter	5	Enterobacter aerogenes profil 1	4
	6	Enterobacter aerogenes profil 2	
	7	Enterobacter cloacae	44
Hafnia	8	Enterobacter hafniae 1	32
	9	Enterobacter hafniae 2	15
Serratia	10	Enterobacter liquefaciens	12
	11	Serratia marcescens	12

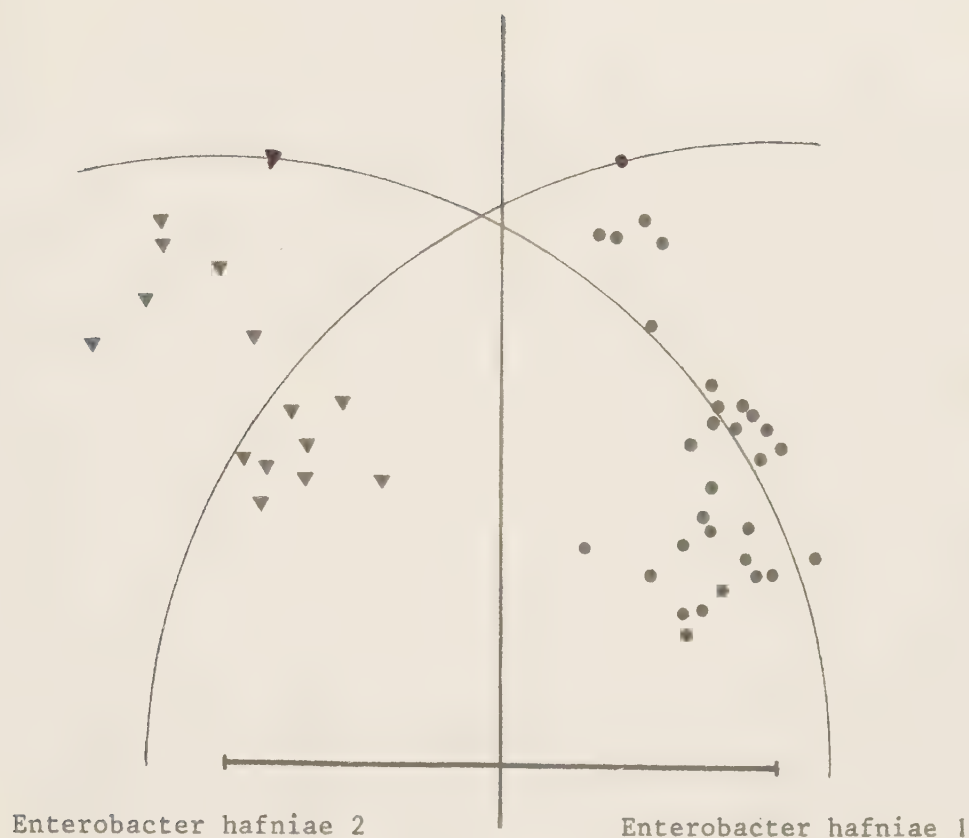


Figure 9 - Schéma montrant les hypersphères et l'hyperplan-séparateur des espèces Enterobacter hafniae 1 et 2

Ce schéma a été obtenu en faisant subir une rotation aux biotypes correspondants aux deux espèces étudiées autour de l'axe passant par les deux spécimens-types correspondant à ces espèces. On peut nettement remarquer que les hypersphères (dont on a dessiné les traces) ne sont pas disjointes, alors que l'hyperplan-séparateur permet de séparer les deux espèces.

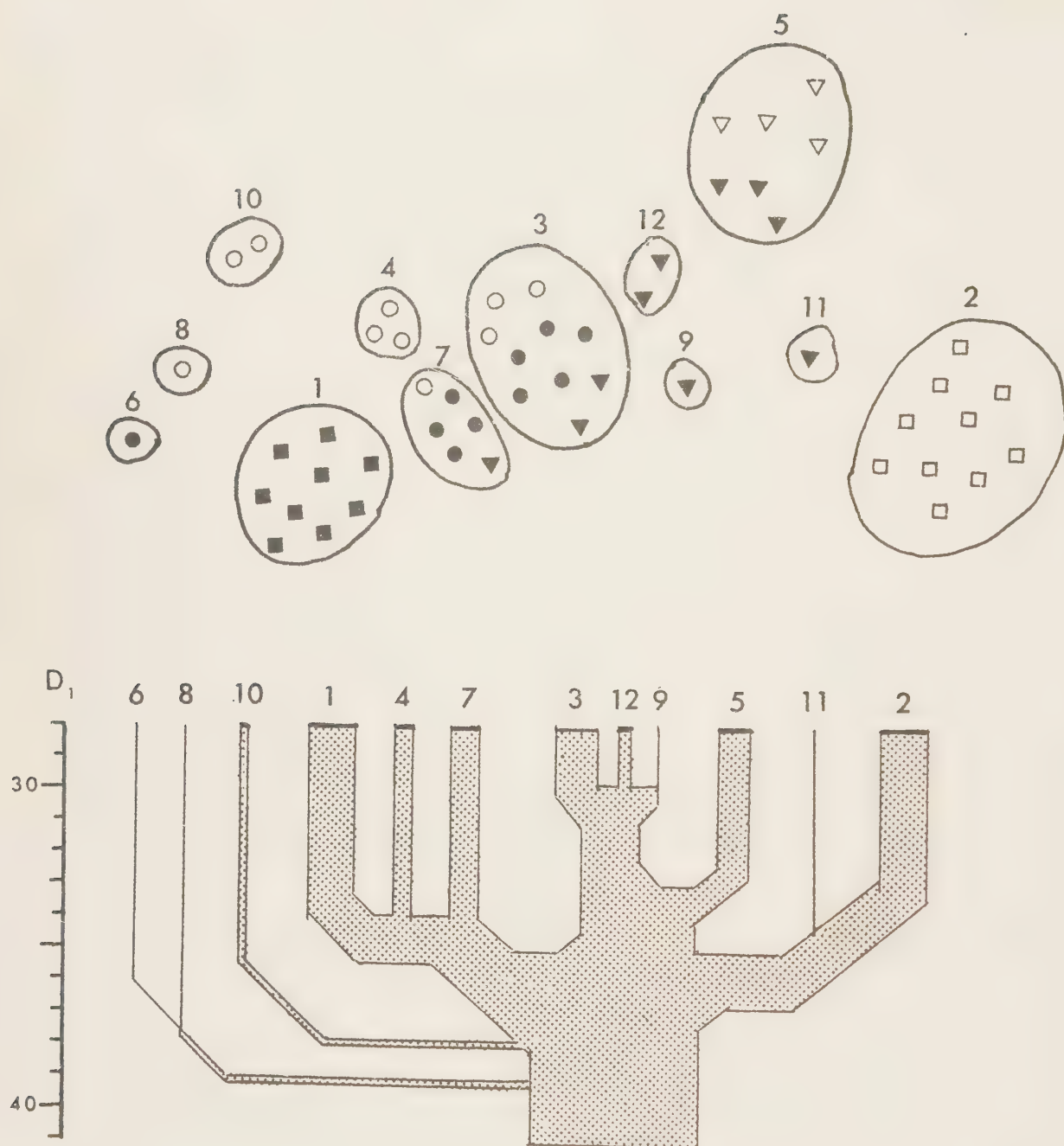


Figure 10 - Etude sur les levures.

Classification de 53 espèces appartenant à 6 genres :

■	Rhodotolura	9
□	Saccharomyces	10
●	Pichia	10
○	Candida	10
▼	Torulopsis	10
▽	Kloeckera	4

évident a priori, car l'existence de deux espèces d'hafnia n'était pas reconnue par tous les microbiologistes ; la figure 9 montre que ces espèces, particulièrement voisines, peuvent donc être différenciées par cette méthode.

L'identification étant le but de ces recherches, les barycentres correspondant à chaque espèce ont été donnés au programme et on lui a ensuite demandé de classer les biotypes. Ce classement, effectué sans erreur, prouve que la notion de groupe centroïde est bien adaptée à la structure de cet échantillon et qu'elle peut être utilisée à des fins d'identification, comme cela sera évoqué dans le paragraphe sur l'aspect système.

3) Travaux effectués sur les levures

Cette étude a été entreprise pour vérifier si les méthodes employées pour les bactéries pouvaient être appliquées aux levures. Du point de vue méthodologique, son principal intérêt était d'essayer un traitement sur des variables qualitatives. Les données ont été prises dans l'ouvrage de LODDER (1971) ; elles correspondent à des caractères morphologiques ou enzymatiques et concernent 53 espèces appartenant à 6 genres. Après codage (effectué selon la méthode décrite au paragraphe II A-3), 67 variables ont été retenues.

La figure 10 montre la composition des groupes obtenus après traitement sur l'ensemble des variables ; la classification est satisfaisante pour les genres Rhodotorula et Saccharomyces, par contre, les genres Pichia, Candida et Torulopsis se répartissent entre plusieurs groupes.

Les résultats obtenus sur des variables morphologiques seules ou sur des variables enzymatiques sont moins satisfaisants.

Compte tenu du fait que l'information utilisée n'avait pas été recueillie en vue d'une exploitation de ce type⁽¹⁾, on peut estimer que cette voie d'approche peut être envisagée pour l'étude des levures ainsi que pour la classification d'échantillons décrits par des variables qualitatives.

(1) Il a été difficile de trouver, dans l'ouvrage considéré, un ensemble de caractères qui soient observés sur la totalité de l'échantillon. Par ailleurs, certains caractères étaient notés "douteux" ce qui a nécessité un codage à trois niveaux : 0 absent - 1 douteux - 2 présent.

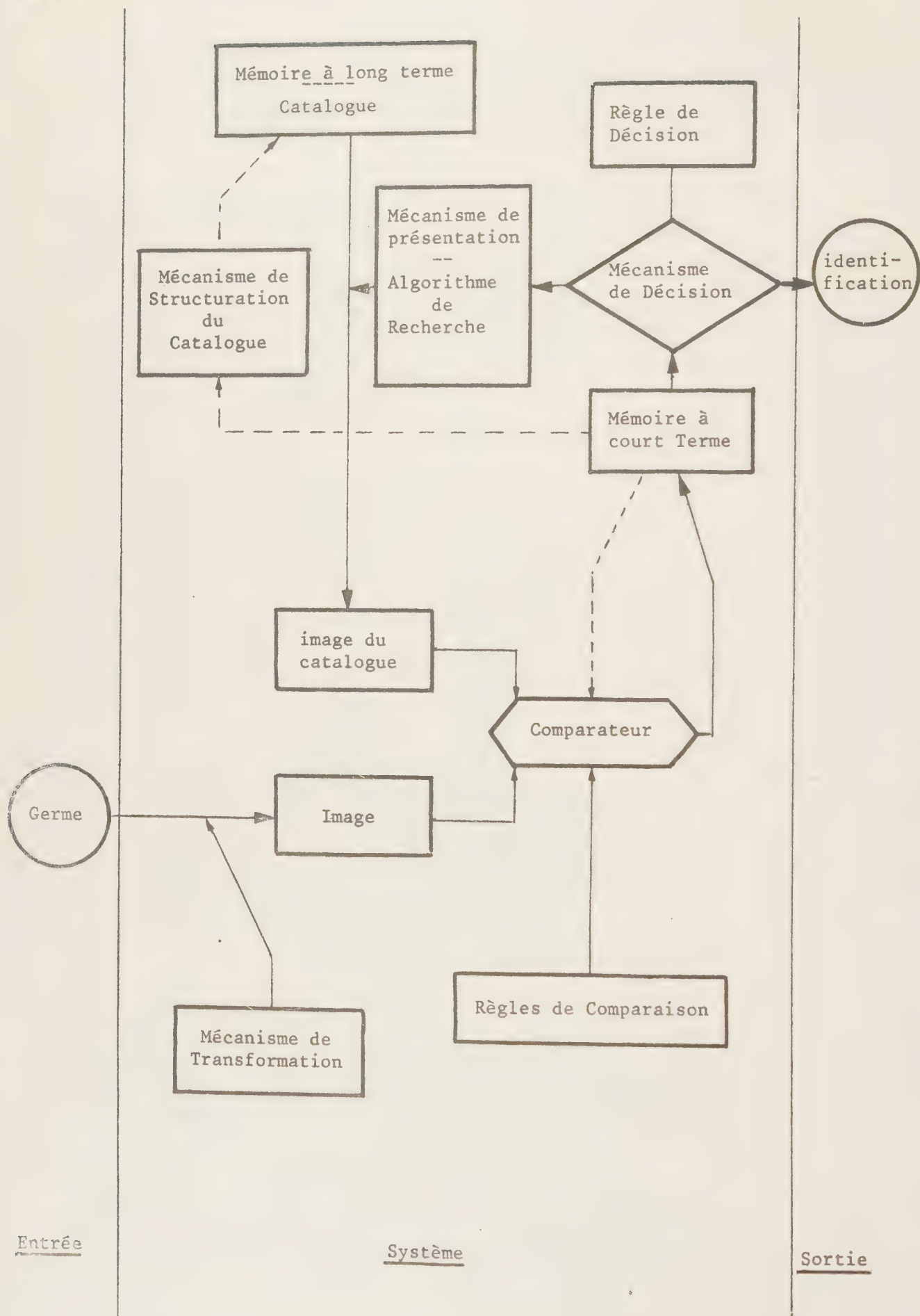


Figure 11 - Représentation sommaire des diverses parties d'un système de reconnaissance des formes.

L'aspect système

Les perspectives de réalisation, dans un proche avenir, d'une machine conçue en vue d'une identification automatisée des micro-organismes, nous ont amené à essayer de définir, en fonction de la méthode retenue, les principales caractéristiques et parties que devrait posséder un tel système.

La figure 11 représente sommairement l'articulation de ces diverses parties : le système reçoit en entrée un germe x non identifié, un mécanisme de transformation (culture, réaction biochimique et lecture) donne une image de ce germe. Cette image est ensuite comparée à une image fournie par un catalogue stocké dans une mémoire à long-terme. Cette comparaison nécessite que soit donné au système des règles de comparaison (critère de la plus courte distance par exemple). Les résultats relatifs à une recherche sont stockés dans une mémoire à court-terme. C'est en fonction de ces résultats et des règles données que le mécanisme de décision procède : soit à la poursuite de la recherche (opération qui met en jeu un algorithme de recherche assurant ainsi la présentation d'une nouvelle image du catalogue), soit de donner en sortie l'identification effectuée.

Cette brève description permet déjà de se rendre compte que la mise au point d'un tel système pose de nombreux problèmes tant en ce qui concerne le mécanisme de transformation (problèmes du type physicochimique) qu'en ce qui concerne le processus de traitement de l'information. La constitution et la structuration de la mémoire à long-terme, ainsi que l'élaboration d'algorithmes optimaux pour la recherche n'en sont encore qu'à une phase conceptuelle. C'est au stade de l'interaction entre la mémoire à court-terme et la mémoire à long terme que se pose le problème de la modification de cette dernière par un apport de connaissances nouvelles (germe nouveau). Dans un premier stade, cette interaction sera probablement effectuée sous le contrôle d'un microbiologiste, mais il semble que l'aptitude d'un système à s'instruire pourrait bien être, dans un proche avenir, une des qualités les plus recherchées.

Conclusion concernant les applications microbiologiques

Les travaux exposés précédemment montrent qu'il existe une bonne concordance entre les études effectuées par les méthodes microbiologiques usuelles et les résultats obtenus par la méthode décrite ici. Cette concordance pourrait faire croire que la méthode est sans intérêt ; il faut toutefois remarquer qu'une étude de ce type représente trente minutes à une heure de calcul

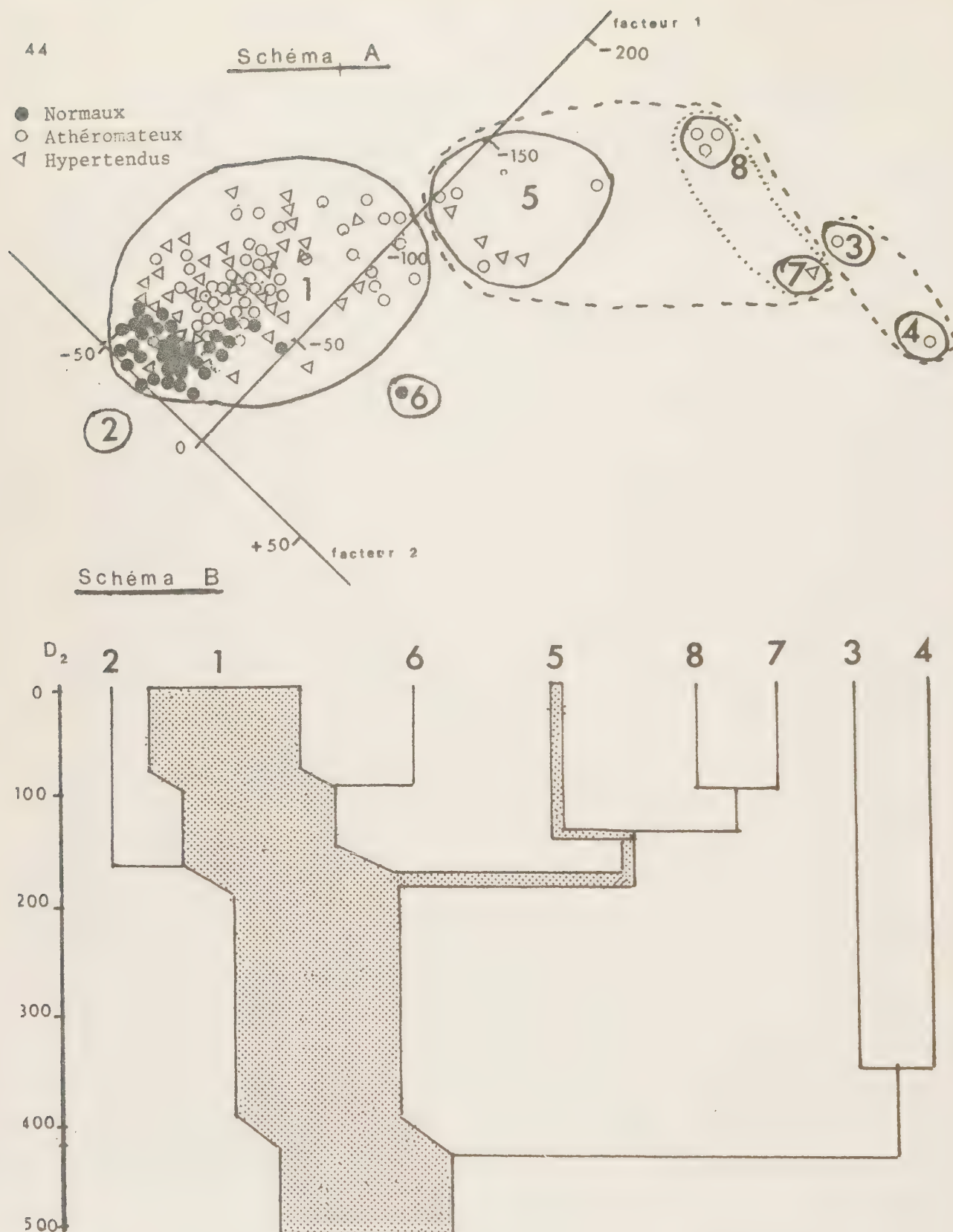


Figure 12 - Classification de 150 sujets en fonction des notes en facteurs calculées à partir de 21 variables biochimiques.

A - Représentation des sujets dans l'espace des deux premiers facteurs.

B - Classification obtenue à partir des cinq premiers facteurs. Les groupes correspondant aux branches du dendrogramme sont représentés sur le schéma A.

sur un ordinateur de capacité moyenne, alors que les techniques usuelles nécessitent le recours à un spécialiste connaissant très bien la population sur laquelle porte l'étude. Cette méthode constitue donc une technique d'appoint non négligeable surtout si on considère l'apport important de données qui seront obtenues par les procédés d'analyse automatique.

B. APPLICATIONS RELEVANT DE LA BIOLOGIE CLINIQUE

Les aspects biochimiques de ces études, entreprises en relation avec le laboratoire de biochimie clinique de l'Hôpital Cardiologique de Lyon, devant faire l'objet d'une thèse à paraître prochainement (ALBRIEUX, 1972), seuls les aspects méthodologiques seront exposés ici.

L'utilisation sur une grande échelle des procédés de dosage automatique en biochimie clinique permet de recueillir, pour chaque sujet qui se présente au laboratoire, une quantité importante d'informations. Par ailleurs, la variabilité des nouvelles mesures, en fonction des divers cas de pathologie, est souvent peu connue et certains syndromes pourraient en fait correspondre à plusieurs types de maladies. Ces deux aspects, qui rendent difficiles l'exploitation des renseignements recueillis, nous ont amené à penser que des méthodes de classification étaient susceptibles d'apporter une aide plus ou moins substantielle au diagnostic. De plus, si on admet qu'une maladie peut être représentée par un certain nombre de sujets avec une configuration particulière de variables, il était logique d'essayer une approche de ce problème par des méthodes de ce type.

Echantillon étudié

Pour cette étude, nous avons choisi un échantillon de 150 sujets sur lesquels 21 variables biochimiques ont été mesurées. Sur la base de critères cliniques, les sujets avaient été classés en trois groupes : normaux, athéromateux et hypertendus) ; mais une des hypothèses de travail était que cette classification pourrait ne pas correspondre à une réalité biologique et qu'il serait possible d'en obtenir une autre à partir des variables mesurées.

Etudes effectuées

Une analyse factorielle globale, portant sur les 150 sujets, a montré que l'échantillon contient beaucoup d'informations puisqu'il faut 12 facteurs pour expliquer 80 % de la variance. Les analyses partielles, effectuées sur chaque groupe, ont montré que les structures des variables

étaient différentes et que le nombre de facteurs explicatifs n'était pas sensiblement diminué.

Ces analyses ont apporté certains renseignements sur l'échantillon étudié, mais ne correspondaient pas exactement au problème posé. Les études entreprises avec le programme Classid ont été effectuées sur les notes en facteurs, calculées à partir de l'analyse factorielle globale et sur les variables standardisées.

1) Etude sur les notes en facteurs

Les raisons du choix des notes en facteurs ont été évoquées à propos des transformations préalables dans la deuxième partie (paragraphe II A-3). Ces notes contiennent l'essentiel de l'information et, par ailleurs, ce traitement élimine une partie de la redondance. Pour les calculs et les représentations graphiques, leur principal avantage est de diminuer les dimensions de l'espace des sujets. De plus, les facteurs étant rangés par ordre d'importance, il était intéressant d'essayer une approche progressive en partant de l'étude de la configuration des sujets dans l'espace à deux dimensions correspondant aux deux premiers facteurs. L'intérêt de cette étude était qu'elle peut être entreprise graphiquement et qu'elle constitue une transition aux études portant sur un plus grand nombre de dimensions (5 à 10 facteurs).

Le schéma 12A représente la configuration des sujets dans l'espace des deux premiers facteurs et montre qu'ils ont tendance à former un groupe plus ou moins compact dont le centre de gravité se situe aux coordonnées -51.52 et -21.97 ; les sujets normaux étant tous situés près de ce centre, légèrement à gauche, alors que les sujets pathologiques s'écartent plus ou moins vers la droite.

Les groupes obtenus à partir d'une classification portant sur cinq facteurs confirment cette tendance : l'arbre de classification du schéma 12B présente un tronc très important sur lequel viennent se greffer des groupuscules.

Les études effectuées en faisant varier les paramètres tels que le nombre de facteurs, le nombre de groupes demandés ou la métrique utilisée ont toujours confirmé cette tendance.

2) Etudes sur les variables standardisées

Compte tenu des remarques du paragraphe II A-3, cette transformation revient à prendre tous les facteurs et à leur donner un poids égal.

Au cours d'une première étude, nous avons demandé 20 groupes et nous avons obtenu 1 groupe de 131 sujets et 19 groupes de 1 sujet. Pour augmenter la discrimination dans le groupe des 131 sujets, un second traitement a été effectué sur ce groupe : nous avons obtenu 1 groupe de 112 sujets et 19 groupes de 1 sujet.

Le fait de prendre tous les facteurs et de leur attribuer un poids égal ne modifie donc pas de manière sensible les résultats.

Interprétation et conclusion

Il semble bien que l'échantillon étudié corresponde à un nuage plus ou moins homogène duquel s'écartent un certain nombre de sujets. Par ailleurs, dans toutes les études effectuées, les sujets très éloignés du centre du groupe étaient toujours les mêmes. Bien que cette constatation aille à l'encontre des hypothèses posées a priori, on aurait en fait un "continuum" qui, à partir des sujets normaux, s'étendrait progressivement aux sujets les plus pathologiques. Dans cette hypothèse, plus un sujet s'écarterait du centre du groupe plus il serait malade.

Sans préjugé des diverses définitions qui peuvent être données du concept de maladie, il est possible de retenir deux modèles qui seront développés à propos de la notion de classe dans la discussion générale.

a) Les sujets atteints d'une maladie forment une classe relativement homogène, bien localisée dans l'espace et nettement séparée des sujets normaux. Certaines maladies infectieuses pourraient appartenir à ce type.

b) Les sujets atteints d'une maladie ne forment pas une classe différente des sujets normaux ; ils s'écartent plus ou moins du groupe formé par ceux-ci sans qu'il soit possible de localiser un groupe homogène spécifique. L'athérome et l'hypertension appartiendraient à cette catégorie.

IV. DISCUSSION GENERALE ET CONCLUSION

La notion de classe est à la base de cette étude et nous avons vu que les définitions résultant de la logique des classes étaient insuffisantes pour permettre l'élaboration des processus de traitement nécessaires à la recherche biologique actuelle. Les différents termes employés par les auteurs qui traitent de ces problèmes (groupe, taxa, "cluster", nuage, etc...) prouvent bien les difficultés inhérentes à la formulation de ce concept, difficultés qui se retrouvent pour la description des configurations obtenues.

Un ensemble d'éléments étant assimilé, par l'analogie géométrique exposée au paragraphe II A-2, à un ensemble de points situés dans un espace à k dimensions, le but d'une méthode de classification est dans ce cas d'étudier la structure de cet ensemble et cela à partir des relations géométriques qui existent entre ses éléments constitutifs. Si les points représentatifs sont répartis d'une manière relativement uniforme dans une certaine région de l'espace, et que le groupe qu'ils forment est séparé des autres points par des distances supérieures à celles qui existent entre eux, on aura une classe homogène. Il n'y a pas de discontinuité dans la représentation spatiale de cette classe et il n'est donc pas possible de séparer, sur de simples considérations géométriques, les éléments qui la constituent sans que cette séparation ne soit arbitraire. Une classe sera hétérogène si les éléments qui la constituent peuvent être regroupés en plusieurs classes homogènes. Ces classes sont alors séparées par des distances plus ou moins importantes et on peut présumer, d'après la représentation géométrique, que leurs éléments constitutifs ont entre eux des relations de similitude qui les différencient globalement de ceux des autres classes. Les espèces biologiques se rapprocheraient des classes homogènes, alors que les genres, les familles et les ordres constituent en général des classes hétérogènes.

Quelques aspects des classes homogènes

Les éléments constitutifs d'une classe homogène peuvent provenir d'un élément unique qui constitue le centre de la classe et que nous appellerons élément-type ou centroïde, c'est le modèle centroïde. Ils peuvent aussi pro-

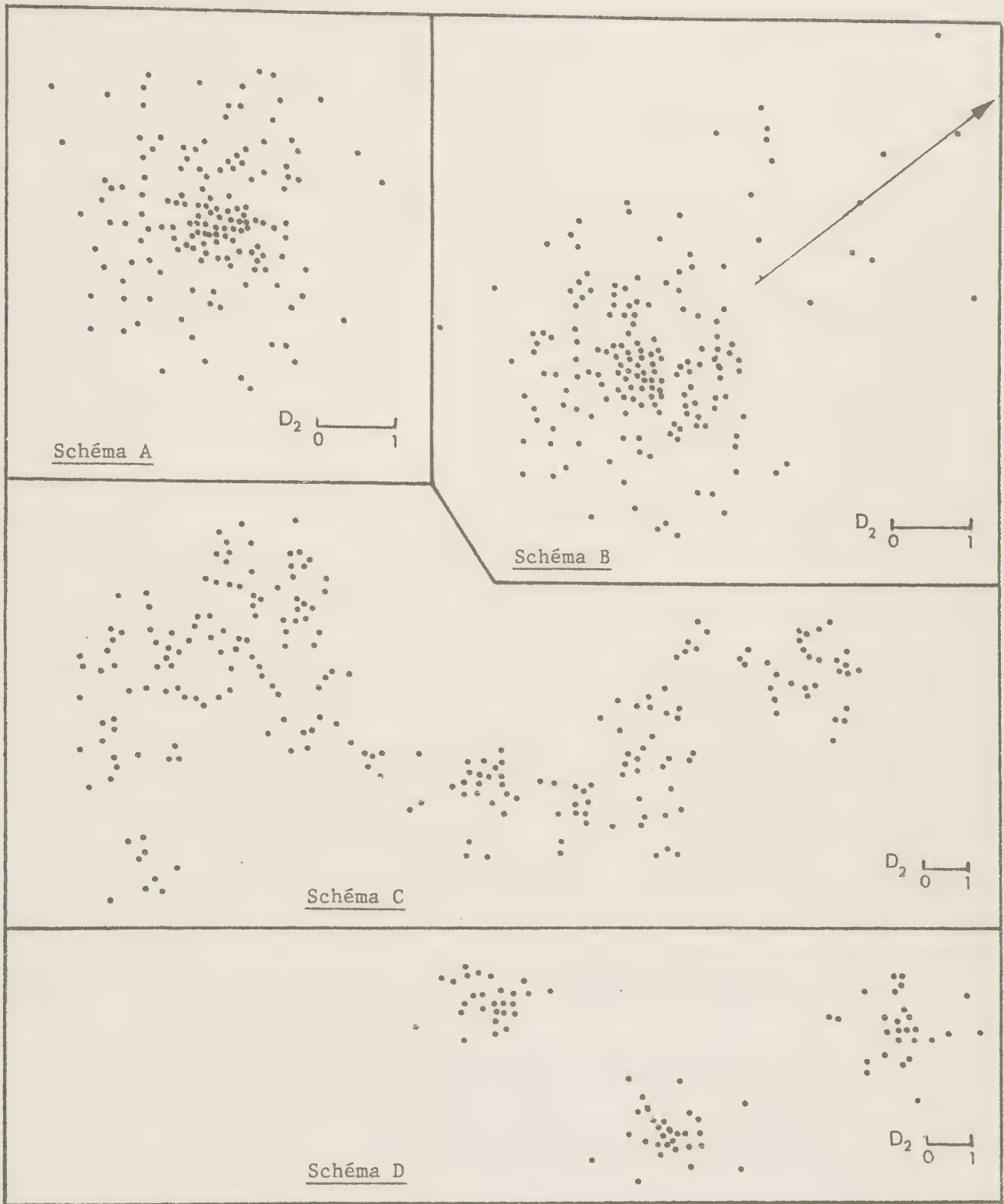


Figure 13 - Représentations de quelques modèles de classes obtenues par simulation.

Les schémas A et B correspondent à des modèles centroïdes; leurs éléments proviennent d'un élément-type. Dans le modèle A la dispersion est uniforme dans toutes les directions; dans le modèle B certains éléments s'écartent du groupe central dans la direction indiquée par la flèche.

Le schéma C représente un modèle à enchaînement.

Le schéma D montre une classe hétérogène constituée par trois classes centroïdes.

venir de plusieurs éléments par un effet de chaîne formant ainsi un modèle à enchaînement.

1) Le modèle centroïde

Dans ce modèle les éléments proviennent d'un élément-type sans qu'ils aient la possibilité de former eux-mêmes d'autres éléments. Dans ces conditions, la variabilité intra-groupe résulte de la dispersion des éléments autour du centroïde, laquelle est fonction de la place disponible ainsi que de la densité (nombre moyen d'éléments par unité de volume). Deux cas peuvent être envisagés :

a) Les variables sont quantitatives

La place disponible à une distance X autour de l'élément-type est fonction du volume de l'hypersphère de rayon X

$$V = K_k X^k$$

pour n pair $K_k = \frac{\pi^{k/2}}{(k/2)!}$

V volume de l'hypersphère de rayon X
k nombre de dimensions de l'espace
 K_k coefficient fonction de k (SOMMERVILLE, 1958)

pour n impair $K_k = \frac{\pi^{(k-1)/2} [(k-1)/2]!}{k!}$

La variation de volume est fonction de la différentielle de V :

$$dV = kK_k X^{k-1} dX$$

La place disponible tend à se déplacer vers les frontières de l'hypersphère lorsque k augmente : si on considère deux sphères de rayons respectifs X_1 et X_2 tel que $X_2 = 2 X_1$ le rapport du volume compris entre les deux hypersphères au volume de la plus petite est :

$$R = \frac{V_2 - V_1}{V_1} = 2^k - 1$$

ce qui donne pour diverses valeurs de n

k	R
2	3
3	7
10	1023
20	1.04 10 ⁶
30	1.07 10 ⁹
40	1.09 10 ¹²
50	1.12 10 ¹⁵

Ces remarques expliquent le fait que les éléments d'un groupe centroïde tendent à s'éloigner de l'élément-type au fur et à mesure que les dimensions de l'espace augmentent.

Il est logique d'admettre que le nombre moyen d'éléments (nombre estimé par la densité) diminue lorsqu'on s'éloigne du centroïde, puisque les éléments proviennent de l'élément-type ; dans l'éventualité contraire le groupe n'aurait pas de frontières et tendrait à occuper de façon homogène tout l'espace disponible.

Si une variable i est normale réduite, la densité de probabilité $y(x_i)$ correspondant à une distance x_i du centre est :

$$y(x_i) = \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2}$$

Si les k variables sont normales réduites et indépendantes, la densité $Y(X)$ correspondant à une distance X du centre de gravité est :

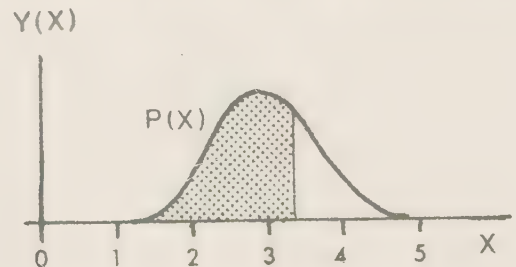
$$Y(X) = \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} e^{-\left(\sum x_i^2\right)/2} = \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} e^{-X^2/2} \quad \text{puisque } X^2 = \sum_{i=1}^k x_i^2$$

Compte tenu des hypothèses, cette densité est constante pour tous les points de l'espace situés à une distance X du centroïde ; le nombre moyen d'éléments à cette distance est donc estimé par :

$$dN_X = K \quad Y(X) \quad dV = K' e^{-X^2/2} X^{k-1} dX$$

Le pourcentage d'éléments dans une hypersphère de rayon X peut être calculé par :

$$P(X) = \frac{\int_0^X e^{-X^2/2} X^{k-1} dX}{\int_0^\infty e^{-X^2/2} X^{k-1} dX}$$



Si N_0 est l'effectif total du groupe, le nombre d'éléments $N(X)$ dans une hypersphère de rayon X est :

$$N(X) = N_0 P(X)$$

Près du centroïde la densité $Y(X)$ est à sa valeur la plus élevée mais il y a peu de place ; lorsqu'on s'éloigne la densité diminue mais le volume augmente, le résultat se traduisant d'abord par une augmentation du nombre moyen d'éléments qui passent par un maximum, lorsque la dérivée (dN_X/dX) s'annule

(c'est-à-dire pour $X = \sqrt{k-1}$), puis diminue au-delà. $P(X)$ est la fonction de répartition d'une loi qui, explicitée par rapport à X^2 , serait une loi de CHI-2 à k degrés de liberté.

Ce modèle est un cas particulier de la loi de LAPLACE-GAUSS dans l'espace à k dimensions (DUGUE, 1958). Il correspond à la situation où la matrice de covariance est égale à la matrice unité (variances égales à 1 et covariances nulles). C'est certainement le modèle le plus simple de classe centroïde; l'ellipsoïde y prend la forme d'une boule. Son principal intérêt est qu'il ne dépend que d'un seul paramètre k , le nombre de dimensions de l'espace où est situé le groupe. En général k est inférieur au nombre des variables descriptives, puisque celles-ci ne sont pas indépendantes : la classe est alors située dans un sous-espace de l'espace défini par les axes correspondant aux variables. Le fait qu'un groupe centroïde soit situé dans un sous-espace ne modifie pas les relations entre les distances et il est possible d'estimer k à partir du maximum de l'histogramme des distances des éléments au centroïde du groupe. Par ailleurs, compte tenu des remarques effectuées dans la deuxième partie, l'utilisation des variables standardisées revient à faire les calculs de distance sur les notes en facteurs pondérées par la racine carrée des racines caractéristiques : on est ainsi dans les conditions d'applications du modèle. La table ci-jointe permet alors d'estimer la probabilité d'appartenance d'un élément à un groupe centroïde.

Bien que les hypothèses requises concernent une situation théorique, qui ne peut être qu'approchée dans les cas réels, les estimations de distances, données par le modèle, sont souvent voisines de celles qui sont calculées à partir des applications.

Lorsque la répartition n'est pas uniforme dans toutes les directions, le groupe peut prendre l'aspect d'une comète ce qui semble être le cas de l'application relevant de la biologie clinique. Les sujets normaux forment la tête de la comète alors que les sujets pathologiques s'étalent dans une certaine direction : ceux qui sont en dehors du groupe central constitué par 112 sujets ont une probabilité d'appartenance à ce groupe inférieure à 1 %.

b) Les variables sont qualitatives

La même démarche peut être suivie pour les variables qualitatives : si les k variables sont dichotomiques les éléments sont répartis sur un treillis de BOOLE et le nombre de positions N_X à une distance X calculée selon la métrique D_1 de MINKOWSKI est :

Valeur du rayon de l'hypersphère en fonction du nombre de dimensions k et de la fonction de répartition.

k	mode	médiane	valeurs correspondant aux probabilités de dépassement $\alpha = 1 - P(X)$		
			0.05	0.01	0.001
1	0.00	0.67	1.96	2.58	3.29
2	1.00	1.17	2.44	3.03	3.61
3	1.41	1.53	2.79	3.36	4.03
4	1.73	1.83	3.08	3.64	4.30
5	2.00	2.08	3.33	3.88	4.53
6	2.24	2.31	3.54	4.10	4.74
7	2.45	2.51	3.75	4.30	4.93
8	2.64	2.70	3.93	4.48	5.11
9	2.83	2.88	4.11	4.65	5.28
10	3.00	3.06	4.27	4.81	5.44
11	3.16	3.21	4.43	4.97	5.59
12	3.31	3.36	4.58	5.12	5.73
13	3.46	3.51	4.73	5.26	5.87
14	3.61	3.65	4.87	5.40	6.01
15	3.74	3.78	5.00	5.53	6.14
16	3.87	3.92	5.13	5.66	6.24
17	4.00	4.04	5.25	5.78	6.39
18	4.12	4.16	5.37	5.90	6.50
19	4.24	4.28	5.49	6.01	6.62
20	4.35	4.40	5.60	6.13	6.73
21	4.47	4.51	5.71	6.24	6.84
22	4.58	4.62	5.82	6.35	6.95
23	4.69	4.73	5.93	6.45	7.05
24	4.79	4.83	6.03	6.55	7.15
25	4.89	4.93	6.13	6.66	7.25
26	5.00	5.03	6.23	6.75	7.35
27	5.09	5.13	6.33	6.85	7.45
28	5.19	5.23	6.43	6.94	7.54
29	5.29	5.32	6.52	7.04	7.63
30	5.38	5.42	6.62	7.13	7.73

$$N_X = \frac{k!}{X!(k-X)!}$$

Lorsque X augmente, N_X croît d'abord jusqu'à atteindre un maximum pour une valeur correspondant à $X = k/2$ ou $X = (k + 1)/2$ (selon que k est pair ou impair) et diminue ensuite.

Si les variables ont une probabilité p de changer et si elles sont indépendantes, l'effectif moyen à une distance X est :

$$E_X = N_X p^X (1-p)^{k-X} = \frac{k!}{X!(k-X)!} p^X (1-p)^{k-X}$$

ce qui correspond à la loi binomiale.

Il est plus difficile d'appliquer ce modèle car les k variables n'ont certainement pas la même probabilité p.

2) Le modèle à enchaînement

Les processus tels que ceux décrits sous le nom de "single linkage" par SOKAL et SNEATH (1963), ou de "l'ultramétrie inférieure maxima" par ROUX (1971), conduisent à la formation de classes à partir d'une relation de similitude ou de proximité : un élément appartient à un groupe s'il peut être relié à ce groupe par une chaîne d'éléments de telle façon que les sauts d'un élément à l'autre soient inférieurs à un certain seuil.

Par l'effet de chaîne qui résulte de ce processus, les éléments d'une même classe peuvent être très dissemblables. Les groupes de ce type ont donc peu de propriétés et on cherchera souvent à obtenir d'autres types de classes. Pourtant de tels groupes peuvent correspondre à une réalité biologique et traduire ainsi un phénomène évolutif, tel que celui de la mutation des micro-organismes, ou un phénomène transitif, comme celui qui peut se produire dans un groupe de communication en sociologie.

La méthode centroïde n'est évidemment pas adaptée à l'étude de ces groupes, mais il est nécessaire de les signaler car on peut incidemment rencontrer une classe de ce type.

Choix de la métrique

Le choix de la métrique est fonction de la relation de ressemblance choisie pour classer les objets : il dépend donc de la nature de l'étude, des caractéristiques des variables retenues et des transformations préalables éventuelles. Il est assez difficile de donner des règles générales : pour les va-

riables quantitatives, il faut souvent tenir compte du facteur d'échelle et des transformations sur les profils s'il y a un phénomène de croissance, par exemple. La distance euclidienne D_2 paraît être une des plus appropriées : le modèle centroïde, lorsqu'il est vérifié, permet alors d'avoir une approximation de l'ordre de grandeur des distances calculées. La distance hypersphérique D_h peut être utilisée lorsque les objets appartenant à une même classe présentent des relations homothétiques. Pour les variables qualitatives, la somme des écarts absolus D_1 devrait permettre d'estimer plus facilement l'effet d'éventuelles pondérations.

Le choix des variables est souvent un cas d'espèce qui se pose au biologiste. Dans les applications exposées précédemment ce choix a été effectué, en relation avec les chercheurs, sur la base de critères empiriques. Il est évident qu'il serait très utile de pouvoir disposer de méthodes permettant de valider des choix qui portent sur un nombre important de variables.

Résultats obtenus en fonction de la nature des classes

La méthode, objet de ce travail, est adaptée à la recherche de classes du type centroïde ; les performances seront bonnes lorsque la structure de l'ensemble sera conforme à ce modèle, comme nous l'avons vu à propos des Klebsiellae.

Dans les autres éventualités, les groupes formés contiennent toujours des éléments semblables ; mais il est évident qu'ils ne reflètent pas la configuration de l'ensemble. L'examen du dendrogramme permet alors de se faire une idée de cette configuration. Lorsque l'ensemble est constitué par une seule classe centroïde, l'arbre est formé par une grosse branche sur laquelle viennent successivement se fixer des branches contenant seulement quelques éléments ; l'application relevant de la biologie clinique était de ce type. Lorsqu'il y a un effet de chaîne, il sera souvent possible de retrouver la transitivité qui existe entre les groupes obtenus ; par ailleurs, si l'ordre d'introduction des éléments est modifié, les groupes formés peuvent aussi être modifiés : cela résulte du fait que la formation de classe centroïde dans un tel ensemble est arbitraire.

Conclusions

La diversité des domaines d'applications et les relations interdisciplinaires que suscitent les problèmes relevant de la formation et de l'identification des classes en biologie ont été seulement esquissées dans ce travail.

Les différents aspects qu'impliquent ces études (distances, processus, etc...) font constamment l'objet de nouveaux développements dont il est souvent difficile de présumer l'intérêt (DAGNELIE, 1966-1968). Dans ce contexte, la simplicité des modèles - essentiellement géométriques -, le souci d'aboutir à des applications concrètes et de dégager des analogies biologiques ont été nos principaux critères.

La validation des résultats obtenus pose de nombreuses difficultés qui ne concernent pas seulement les classifications élaborées par des processus mathématiques mais également celles qui résultent des autres méthodes. A défaut de critères objectifs, nous avons essayé de juger les résultats par recoupement avec ceux obtenus par d'autres techniques généralement empiriques. Là encore, le problème est complexe car les chercheurs qui demandent des études de ce type contestent généralement plus ou moins les classifications usuelles. Enfin, il est admis, comme le rappelle SNEATH (1970), que l'absence de tests rigoureux pour estimer la présence de groupes est une déficience grave, déficience sur laquelle vient se greffer l'absence de différents modèles de classes. En général, le biologiste recherche des classes centroyées. Il est souvent désappointé de ne pas voir ses hypothèses se confirmer et, dans cette éventualité, il lui est également difficile de décrire les résultats obtenus. Nous pensons que ces deux aspects sont liés et que l'élaboration de modèles de classes, précédemment évoquée, devrait permettre la recherche de tests pour l'estimation de la présence de classes d'un type donné. Il est nécessaire que ces étapes, qui font partie des études relatives à la formation des classes, trouvent une solution satisfaisante pour que l'identification puisse être envisagée.

B I B L I O G R A P H I E

- ALBRIEUX M.J. - 1972 - Etude du profil biochimique multiparamétrique de l'athérosclérose.
Thèse d'Etat en Pharmacie, LYON. (à paraître).
- ANDERSON M. - 1971 - L'ethnobiologie - Les classifications populaires.
La Recherche, 18, 1031-1039.
- ATKINSON A.C. - 1969 - Models for memory. Dans les Actes du colloque OTAN sur "La recherche en enseignement programmé".
Dunod, Paris, 75-120.
- BARBE J. - 1969 - Organisation méthodique de l'étude des caractères enzymatiques des bactéries de la tribu des Klebsiellaeae : Application à la classification.
Thèse d'Etat, Faculté mixte de Médecine et de Pharmacie de Marseille. 247 p.
- BELIS M. - 1966 - Les systèmes éducatifs.
Cybernetica, 9, 157-167.
- BOYCE A.J. - 1969 - Mapping diversity : a comparative study of some numerical methods. In Numerical taxonomy, A.J. COLE.
Academic Press, London and New-York, 1-31.
- BUISSIÈRE J. et COLOBERT L. - 1969 - Normalisation des épreuves destinées à la mise en évidence de l'équipement enzymatique des microorganismes. Les procédés "univariants".
Path. Biol., 17, 301-312.
- BUISSIÈRE J. et NARDON P. - 1968 - Microméthode d'identification des bactéries. - I - Intérêt de la quantification des caractères biochimiques.
Ann. Inst. Pasteur, 115, 218-231.
- COUFFIGNAL L. - 1968 - La cybernétique.
P.U.F., Paris.
- DAGNELIE P. - 1966 - A propos des différentes méthodes de classification numérique.
Rev. Stat. Appl., 14, 55-75.
- DAGNELIE P. - 1968 - Introduction aux problèmes et aux méthodes de classification numérique.
Biom. Prax. 9, 8-111.

- DEFAYOLLE M., COLOBERT L., PONCET P., BUISSIÈRE J. et PONTIER J. - 1968 -
Application de l'analyse factorielle à la taxonomie des
microorganismes.
Biom. Praxis, 9, 14-51.
- DOZOL F. - 1957 - L'analyse configurationnelle.
Bull. du Centre d'Et. et Rech. Psychot., 7, 469-483.
- DUCIMETIÈRE P. - 1970 - Les méthodes de la classification numérique.
Rev. Stat. Appl., 18, 5-25.
- DUGUE D. - 1958 - Traité de statistiques théoriques.
Masson, Paris, 313 p.
- GERARDIN L. - 1968 - La Bionique.
Hachette, Paris, 254 p.
- GRAUFE D. and LYNN W.J. - 1969 - Some aspects regarding mechanistic
modelling of recognition and memory.
Cybernetica, 3, 119-141.
- GRIZE J.B. - 1967 - La logique des classes.
Dans Logique et connaissance scientifique, J. PIAGET.
Gallimard, Paris, 143-173.
- JACQ J., DEFAYOLLE M. et BUISSIÈRE J. - 1971 - Modèles et système
d'identification des bactéries.
Dans les Conférences des Journées d'Informatique médicale.
Inst. de Rech. d'Informatique et d'Automatisme, Rocquencourt,
25-38.
- JARDINE N. - 1971 - A new approach to pattern recognition.
Nature, 234, 526-528.
- KRUSKAL J.B. - 1964 - Nonmetric multidimensional scaling,
A numerical method.
Psychometrika, 29, 115-129.
- LANGE G.N., WILLIAM W.T. - 1967 - A general theory of classificatory
sorting strategies. 1 - Hierarchical systems.
Computer Journal, 9, 373-380.
- LODDER J. and KREGER-VAN RIJ N.J.W. - 1971 - The Yeast, a taxonomic study.
North Holland Publishing Company, Amsterdam, 713 p.
- PARK J.M. - 1969 - Classification of mixed-mode data by R-mode factor
analysis and Q-mode cluster analysis on distance fonction.
In Numerical taxonomy, A.J. COLE.
Academic Press, London and New-York, 216-223.
- PAULE R. - 1971 - Contribution à l'étude du genre *Lactobacillus*
par une méthode normalisée.
Thèse d'Université en Pharmacie, Lyon, 159 p.
- RICARD M. - 1971 - Les migrations animales.
Laffont, Paris, 252 p.

- SNEATH P.H.A. - 1969 - Evaluation of clustering methods.
In Numerical taxonomy, A.J. COLE.
Academic Press, London and New-York, 257-271.
- SOKAL R.R. - 1966 - Numerical taxonomy.
In Mathematical thinking in behavioral sciences, 1968
W.H. FREEMAN and C°, San Francisco, 174-184.
- SOKAL R.R. and SNEATH P.H.A. - 1963 - Principles of numerical taxonomy.
W.H. FREEMAN and C°, San Francisco and London, 359 p.
- SOMMERVILLE D.H.Y. - 1958 - An introduction to the geometry of N dimensions.
Dover publications, Inc., New-York.
- THEODORIDES J. - 1965 - Histoire de la biologie.
P.U.F., Paris, 124 p.
- TINBERGEN N. - 1969 - La vie sociale des animaux.
Payot, Paris, 186 p.
- WIENER N. - 1948 - Cybernetics, or control and communication in the
animal and the machine.
Wiley, New-York.
- WRIGHT O.S., TAYLOR A., ROY DAVIES D., LEE S.G.M. and REASON J.T. - 1970 -
Remembering. In Introducing psychology - An experimental
approach.
Penguin Book, Ltd. Harmondsworth, Middlesex, England. 345-376.

PROGRAMME D'ETUDE DES CLASSIFICATIONS

EN VUE DE L'IDENTIFICATION

```

*****  **          **  *****  *****  *****  *****
*****  **          ***  *****  *****  *****  *****
**      **          **   *      *      *      *      *      *
**      **          ****  ****  ****  ****  ****  ****  ****
**      **          ****  ****  ****  ****  ****  ****  ****
**      **          ****  ****  ****  ****  ****  ****  ****
**      **          ****  ****  ****  ****  ****  ****  ****
*****  *****  ****  ****  ****  ****  ****  ****  ****
*****  *****  ****  ****  ****  ****  ****  ****  ****

```

Ce programme a été mis au point sur un ordinateur CII 10 020 (24 K). Avec ses dimensions actuelles, son encombrement mémoire est de 12 986 mots de 16 bits.

```

COMMON B(1270),KZ(10),Z(10)
COMMON KG(200),IDENT(200),NG(30),C(19),CBEF(70)
CALL TRAIT (B,KZ,Z,KG,IDENT,NG,C,CBEF)
STOP
END

```

```

C          SUBROUTINE TRAIT
SUBROUTINE TRAIT (B,KZ,Z,KG,IDENT,NG,C,CBEF)
  DIMENSION B(1),KZ(1),D(19),F(19),G(19),H(19),Z(1),KG(1),
  IDENT(1),NG(1),C(1),CBEF(1)
  DIMENSION DX(30),IA(30),IB(30)
  NET=0
  WRITE(6,190)
190 FORMAT(1H1,23HPROGRAMME C L A S S I D//)
  READ (5,191)KAR,(C(I),I=1,19)
  WRITE(6,191)KAR,(C(I),I=1,19)
  L FORMAT(1X,I2,19A4)
  READ (5,191)KKD,D,KKF,F,KKG,G,KKH,H
  WRITE(6,191)KKD,D,KKF,F,KKG,G,KKH,H
  READ (5,193)KAR,KM,LC,LM
  WRITE(6,193)KAR,KM,LC,LM
  READ (5,193)KAR,IZM,IFOR,KDIV,KMUL
  WRITE(6,193)KAR,IZM,IFOR,KDIV,KMUL
  READ (5,193)KAR,ITRAN,KINT
  WRITE(6,193)KAR,ITRAN,KINT
  FORMAT(1X,I2,10I5)

  ITRAN=ITRAN+1
  GO TO (90,90,91),ITRAN
  IDPL=1
  IF IDPL=1
  IDPL=2

C  92 IF(KDIV)102,102,103
C  92 DO 108 I=1,IZM
100 CBEF(I)=1.
  GO TO 109
103 READ(5,D)(CBEF(I),I=1,IZM)
  WRITE(6,D)(CBEF(I),I=1,IZM)
105 IF(KMUL)112,112,110
110 READ(5,F)(B(I),I=1,IZM)
  WRITE(6,F)(B(I),I=1,IZM)
  DO 111 IZ=1,IZM
111 CBEF(IZ)=CBEF(IZ)/B(IZ)

C  112 WRITE(6,910)
910 FORMAT(1H1,23X,21HFORMATION DES GROUPES///
  13X,22HUSION DE DEUX GROUPES,19X,23HASSIGNATION A UN GROUPE///
  35X,5HECART,5X,6HGRUPE,4X,6HGRUPE,2X,4HEFF.,3X,1H*,4X,5HECART,
  44X,8HINDIVIDU,3X,6HGRUPE,2X,4HEFF.,3X,1H*//)
  KGR=KM-1
  NA0=KM*IZM
  NA1=KM*IZM+1
  NA2=NA1+IZM+1

C          PRISE D INFORMATION INITIALE

  DO 203 K=1,KM
  KG(K)=K

```

```

      NG(K)=1
      NP1=(K-1)*IZM+1
      NP2=NP1+IZM-1
      READ(5,5) ILENT(K), (B(NB), NB=NB1, NB2)
      WRITE(6,205) IDENT(K), K, NG(K)
205  FORMAT(40X, 1H*, 14X, I6, 4H ->, I4, 2X, I5, 4X, 1H*)
      GO TO (203, 203, 206), IFBR
206  CALL NORMA (IZM, B, K, CBEF)
      WRITE(6,9000) (B(I), I=NP1, NB2)
9000  FORMAT(1X, 10F7.4)
203  CONTINUE
      IF(LM-KM) 99, 395, 204
204  LP=KM+1
C
C      PROCESSUS DE CLASSIFICATION
C
104  CALL CALPRE ( 1, B, KZ, H, Z, KG, IDENT, NG, CBEF,
1KM, LP, IZM, IFBR, KDIV, KMPL, NAD, NA1, NA2, LP, IDPL, D12, DP3, K1, K2, K3, KA)
      CALL CALPRE ( 2, B, KZ, H, Z, KG, IDENT, NG, CBEF,
1KM, LP, IZM, IFBR, KDIV, KMPL, NAD, NA1, NA2, LP, IDPL, D12, DP3, K1, K2, K3, KA)
C
      WRITE(6,900) IDENT(LP), K3, DP3, K2, K1, D12
900  FORMAT(1H-, 30X, 2I5, F10.3, 2I5, F10.3)
      IF(D12-DP3) 350, 360, 360
C
C      FUSION DE DEUX GROUPES
C
350  II=LP-1
      DO 351 I=1, II
      IF(KG(I)-K2) 351, 352, 351
352  KG(I)=K1
351  CONTINUE
C
      CALL BARY (IFBR, IZM, B, NG, K1, K2, CBEF, DIST, DEPL)
      NZD=(K2-1)*IZM
      DO 353 IZ=1, IZM
      I=NZD+IZ
      J=AD+IZ
353  B(I)=B(J)
      WRITE(6,354) D12, K2, K1, NG(K1)
354  FORMAT(1H-, F10.3, 4X, I6, 4H ->, I4, 2X, I5)
C
      NG(K2)=1
      KG(LP)=K2
      WRITE(6,205) IDENT(LP), K2, NG(K2)
      GO TO 390
C
C      ASSIGNATION DE P A UN GROUPE
C
360  KG(LP)=K3
      CALL BARY (IFBR, IZM, B, NG, K3, KA, CBEF, DIST, DEPL)
      WRITE(6,1390) DP3, IDENT(LP), K3, NG(K3)
1390  FORMAT(40X, 1H*, F10.3, 4X, I6, 4H ->, I4, 2X, I5, 4X, 1H*)
390  CONTINUE
      LP=LP+1
      IF(LP-LC) 104, 104, 370
370  IF(LP-LC) 470, 470, 469
C

```

PROCESSUS D IDENTIFICATION

```

468 WRITE(6,468)
468 FORMAT(1X/14HIDENTIFICATION)
471 CALL CALPR8 ( 2,B,KZ,H,Z,KG,IDENT,NG,COEF,
1KM,LM,IZM,IFOR,KDIV,KMUL,NAD,NA1,NA2,LP,IDPL, D12,DP3,K1,K2,K3,KA)
WRITE(6,1390)DP3,IDENT(LP),K3,NG(K3)
LP=LP+1
IF(LP-LM)471,471,470

```

IMPRESSION DES GROUPES

```

470 WRITE(6,911)
911 FORMAT(1H1///20X,22HIMPRESSION DES GROUPES//
13X,6HGRUPE,6X,9HINDIVIDUS///)
MMM=10
M=0
DO 371 K=1,KM
WRITE(6,916)
916 FORMAT(1X)
DO 371 L=1,LM
IF(KG(L)-K)372,373,372
373 M=M+1
KZ(M)=IDENT(L)
372 IF(M-MMM)374,375,99
374 IF(L-LM)371,377,99
377 IF(M)371,371,375
375 WRITE(6,376)K,(KZ(I),I=1,M)
376 FORMAT(1X,16,5X,10I6)
M=0
371 CONTINUE

```

IMPRESSION DES BARYCENTRES

```

WRITE(6,914)
FORMAT(1H1///20X,23HDESCRIPTION DES GROUPES/
11X,6HGRUPE,3X,4HEFF.///)
KA=KM+1
DO 381 K=1,KM
CALL CENTRE (IFOR,IZM,B,NG,K,KA)
WRITE(6,383)K,NG(K),(B(IZ),IZ=NA1,NA2)
383 FORMAT(1X,14,18,10F10.3/(13X,10F10.3))
WRITE(6,916)
381 CONTINUE

```

CALCUL ET IMPRESSION DES DISTANCES INTERGROUPES

```

395 WRITE(6,913)
913 FORMAT(1H1)
GO TO (397,398,398),ITRAN
397 IDPL=1
GO TO 399
398 IDPL=2
399 CALL CALPR8 ( 3,B,KZ,H,Z,KG,IDENT,NG,COEF,
1KM,LM,IZM,IFOR,KDIV,KMUL,NAD,NA1,NA2,LP,IDPL, D12,DP3,K1,K2,K3,KA)

```

PROCESSUS DE STRUCTURATION - REGROUPEMENT

```

400 CONTINUE
   IF(KM-KGR=1)99,403,404
403 WRITE(6,912)IDPL
912 FORMAT(1H1///20X,12HREGRGROUPEMENT,10X,7HCRITERE,3X,13///)
   KTRAC=KINT+1
   GO TO (401,402),KTRAC
401 KTRAI=4
   GO TO 410
402 KTRAI=3
404 WRITE(6,918)
918 FORMAT(1X//13H REGRGROUPEMENT//3H N0,2X,5HECART,
  17X,4HDANS,12X,11HDESCRIPTION//)
410 NET=NET+1
   DX(NET)=D12
   IA(NET)=K1
   IB(NET)=K2
   WRITE(6,1410)NET,D12,K2,K1
1410 FORMAT(1H-,I2,F8.3,2X,I4,4H -> ,I4)
   CALL BARY (IF0R,IZM,B,NG,K1,K2,C0EF,DIST,DEPL)
   KA=KM+1
   CALL CENTRE (IF0R,IZM,B,NG,K1,KA)
   WRITE(6,1005)NG(K1),(B(IZ),IZ=NA1,NA2)
1005 FORMAT(1X,24X,I4,10F10.3/(29X,10F10.3))
   KGR=KGR+1
   IF(KGR=1)420,419,419
419 CALL CALPR8 (KTRAI,B,KZ,H,Z,KG,IDENT,NG,C0EF,
  1KM,LM,IZM,IF0R,KDIV,KMUL,NAD,NA1,NA2,LP,IDPL, D12,DP3,K1,K2,K3,KA)
   GO TO 400
420 CALL ARBRE (KM,DX,IA,IB)
   GO TO 341
99 WRITE(6,179)
179 FORMAT(1X,25H8RTIE ANORMALE SUR UN IF/)
341 CONTINUE
   RETURN
   END

```

```

SUBROUTINE CALPRO (KTRAI,B,KZ,H,Z,KG,IDENT,NG,CBEF,
1KM,LM,IZM,IFOR,KDIV,KMUL,NAD,NA1,NA2,LP,IDPL, D12,DP3,K1,K2,K3,KA)
DIMENSION B(1),KZ(1),H(1),Z(1),KG(1),IDENT(1),CBEF(1),NG(1)
      VARIATION SUR LA LIGNE L= LD,LF
      GO TO (105,106,105,105),KTRAI
105 LD=2
      LF=KM
      MAT=0
      MATI=MAT
      GO TO 107
106 LD=1
      LF=1
107 L=LD
      VARIATION SUR LES OBJETS LP=(KM+1),KM
110 KD=1
      GO TO (111,112,111,111),KTRAI
111 KF=L-1
      KA=L
      NG(KA)=NG(L)
      GO TO 114
112 KF=KM
      KA=KM+1
      READ(5,H)IDENT(LP),(B(I),I=NA1,NA2)
      NG(KA)=1
      GO TO (114,114,214),IFOR
214 CALL NORMA (IZM,B,KA,CBEF)
      WRITE(6,9000)(B(I),I=NA1,NA2)
9000 FORMAT(1X,10F7.4)
114 CONTINUE
      M=0
      MT=1
      K=KD
      VARIATION SUR LA COLONNE K=KD,KF
      CALCUL D UNE LIGNE
120 CONTINUE
      GO TO (224,224,221,221),KTRAI
221 IF(NG(L))131,131,222
222 IF(NG(K))314,314,224
224 M=M+1
      MAT=MAT+1
      MATI=MAT
      CALL PROXI (IFOR,IZM,B,NG,K,KA,CBEF,DIST,DEPL)
      IF(IDPL-1)227,227,228
227 W=DIST
      GO TO 229
228 W=DEPL
229 CONTINUE
      301 RECHERCHE DISTANCE MINIMALE INTER-GROUPE
      D12 ENTRE GROUPE K1 ET K2 KTRAI=-1
      GO TO (301,303,301,301),KTRAI
301 IF(MAT-1)311,311,310
311 D12=W
      K1=K

```

```

      K2=L
      GO TO (125,119,302,125),KTRAI
310 IF(D12=W)314,314,313
313 D12=W
      K1=K
      K2=L
314 GO TO (125,119,302,125),KTRAI
                                302  STOCKAGE ET IMPRESSION DISTANCE INTERG
302 Z(M)=DIST
      IF(MAT1=1)315,315,317
315 WRITE(6,316)
316 FORMAT(1X///23H DISTANCES INTERGROUPES///)
      MAT1=MAT1+1
      I1=0
      DO 304 I=1,K
      IF(NG(I))304,304,305
305 I1=I1+1
      IDENT(I1)=I
304 CONTINUE
      I1=I1-1
      WRITE(6,304)(IDENT(I),I=1,I1)
306 FORMAT(1X,7X,10I10)
      WRITE(6,915)
915 FORMAT(1X//)
317 IF(M=1)125,121,122
121 LIG=L
122 IF(MT=KF)123,124,124
123 IF(M=10)125,124,124
124 WRITE(6,307)LIG,(Z(I),I=1,M)
307 FORMAT(1X,16,4X,10F10.3)
      M=0
      GO TO 125
                                303  RECHERCHE DISTANCE MINIMALE DP3 ENTRE
                                L OBJET P ET LES GROUPES LP3
303 IF(K=1)330,331,330
331 DP3=W
      K3=K
      GO TO 125
330 IF(OP3=W)125,125,332
332 LP3=W
      K3=K
      GO TO 125
125 K=K+1
      MT=MT+1
      IF(K=KF)120,120,131
                                131  L=L+1
                                IF(L=LF)111,110,119
119 RETURN
      END

```

```

SUBROUTINE PR9XI
SUBROUTINE PR9XI (IFR, IZM, B, NG, KB, KA, COEF, DIST, DEPL)
DIMENSION P(1), NG(1), COEF(1)
SX=0.
SY=0.
SX2=0.
SY2=0.
SXY=0.
DIF=0.

                                VARIATION SUR IZ=1, IZM
                                CALCUL D'UN ELEMENT

DO 130 IZ=1, IZM
  I=(KB-1)*IZM+IZ
  J=(KA-1)*IZM+IZ
  GO TO (131, 131, 132), IFR
131 CONTINUE
  X=P(I)/FLOAT(NG(KB))/COEF(IZ)
  Y=P(J)/FLOAT(NG(KA))/COEF(IZ)
  GO TO 133
132 X=P(I)
  Y=P(J)
133 SX=SX+X
  SY=SY+Y
  SX2=SX2+X*X
  SY2=SY2+Y*Y
  SXY=SXY+X*Y
  DIF=DIF+ABS(X-Y)
130 CONTINUE

                                CHOIX DE LA FORMULE
  GO TO(1, 2, 3), IFR
  1 DIST=DIF
  GO TO 150
  2 DIST=SQRT(SX2+SY2-2.*SXY)
  GO TO 150
  3 DIST=ATAN(SQRT(1.-SXY**2)/SXY)
150 CONTINUE
  XPEN=FLOAT(NG(KA)*NG(KB))/FLOAT(NG(KA)+NG(KB))*2.
  DEPL=DIST*XPEN
  RETURN
END

```

```

C SUBROUTINE NORMA
SUBROUTINE NORMA (IZM,B,K,CBEF)
DIMENSION B(1),CBEF(1)
ANBRM=0.
DO 100 IZ=1,IZM
  I=(K-1)*IZM+IZ
  B(I)=B(I)/CBEF(IZ)
100 ANBRM=ANBRM+B(I)**2
  ANBRM=SQRT(ANBRM)
DO 110 IZ=1,IZM
  I=(K-1)*IZM+IZ
110 B(I)=B(I)/ANBRM
RETURN
END

```

```

C SUBROUTINE CENTRE
SUBROUTINE CENTRE (IFOR,IZM,B,NG,K,KA)
DIMENSION B(1),NG(1)
GB TO (1,1,3),IFOR
1 DO 100 IZ=1,IZM
  I=(KA-1)*IZM+IZ
  J=(K-1)*IZM+IZ
100 B(I)=B(J)/FLOAT(NG(K))
  GB TO 500
3 DO 110 IZ=1,IZM
  I=(KA-1)*IZM+IZ
  J=(K-1)*IZM+IZ
110 B(I)=B(J)
500 RETURN
END

```

```

SUBROUTINE BARY
SUBROUTINE BARY (IFBR, IZM, E, NG, K1, K2, CBEF, DIST, DEPL)
DIMENSION B(1), NG(1), CBEF(1)
XN1=FLOAT(NG(K1))
XN2=FLOAT(NG(K2))
GO TO (1,1,3), IFBR
1  DO 100 IZ=1, IZM
    I=(K1-1)*IZM+IZ
    J=(K2-1)*IZM+IZ
100  B(I)=B(I)+E(J)
    NG(K1)=NG(K1)+NG(K2)
    NG(K2)=0
    GO TO 500
3  CALL PR0XI (IFBR, IZM, B, NG, K1, K2, CBEF, DIST, DEPL)
    ABB=DIST
    GD=COS((ABB)/2.)
    ABG=ABB*XN2/(XN1+XN2)
    EBG=ABG-ABF/2.
    DP=BD*SIN(EBG)/COS(EBG)
    DIF=0.
    DO 101 IZ=1, IZM
        J=(K2-1)*IZM+IZ
        I=(K1-1)*IZM+IZ
101  DIF=(B(I)-B(J))*2+DIF
        AB=SQRT(DIF)
        PA=AB/2.+DP
        PB=AB-PA
        ANORM=0.
        DO 300 IZ=1, IZM
            I=(K1-1)*IZM+IZ
            J=(K2-1)*IZM+IZ
            B(I)=(B(I)*PB+E(J)*PA)/AB
300  ANORM=ANORM+B(I)**2
        ANORM=SQRT(ANORM)
        DO 310 IZ=1, IZM
            I=(K1-1)*IZM+IZ
310  B(I)=B(I)/ANORM
        NG(K1)=NG(K1)+NG(K2)
        NG(K2)=0
500  RETURN
END

```

```

SUBROUTINE ARBRE (KM,DX,IA,IB)
DIMENSION X(1),IA(1),IB(1)
DIMENSION KAR(30),KBR(30),KIR(30),KUR(30),MEM(30)
DIMENSION IA1(30)
DATA KZE/' '/,KAST/'*'/
ELABORATION D UN ARBRE A KM RAMEAUX A PARTIR DE IB -> IA
LM=KM-1
DO 30 L=1,LM
  KBR(L)=KAST
30 IA1(L)=IA(L)
  KBR(KM)=KAST
  K=1
  KAR(K)=1
  M=1
  MEM(M)=1

      CALCUL DU VECTEUR DE BASE (KAR(K),K=1,KM)
50 DO 100 L=1,LM
  IF(IA(L)-MEM(M))100,101,100
101 IA(L)=0
  K=K+1
  KAR(K)=IB(L)
  M=M+1
  MEM(M)=IB(L)
  L=0
100 CONTINUE
  MEM(M)=0
  M=M-1
  IF(M)102,102,50
102 WRITE(6,900)(KAR(K),K=1,KM)
900 FORMAT(1H1,23HAKERE DE CLASSIFICATION,10X,12HDENDROGRAMME
1//28X,30I3)
  WRITE(6,902)(KBR(L),L=1,KM)
  WRITE(6,902)(KBR(L),L=1,KM)
  WRITE(6,902)(KBR(L),L=1,KM)

      COMPOSITION ET IMPRESSION D UNE LIGNE
DO 200 L=1,LM
  KA=IA1(L)
  KB=IB(L)
DO 250 K=1,KM
  KBR(K)=KZE
  KUR(K)=KZE
  KIR(K)=KZE
  IF(KAR(K)-IA1(L))250,251,250
251 KA=K
250 CONTINUE
DO 260 K=1,KM
  IF(KAR(K)-IB(L))260,261,260
261 KB=K
260 CONTINUE
DO 300 K=1,KM
  IF(KAR(K))300,300,302
302 KBR(K)=KAST
  KUR(K)=KAST

```

```

300 CONTINUE
   KUR(KB)=KZE
   DO 400 K=KA,KB
     KBR(K)=KAST
400   KIR(K)=KAST
     KIR(KB)=KZE
     KAR(K)=0
     WRITE(6,901)L,DX(L),IA(L),IA1(L),(KBR(K),KIR(K),KIR(K),K=1,KM)
901   FORMAT(1X,I3,F8.3,I7,4A1->,I3,5X,90A1)
     WRITE(6,902)(KUR(K),K=1,KM)
902   FORMAT(1X,20X,30(A1,2X))
200 CONTINUE
   RETURN
   END

```

Vu.

Les membres du jury :

MM. LEGAY

Président

TEULADE

LAZAR

DEFAYOLLE

Examineurs

Vu et approuvé.

Le Directeur de l'U.E.R.
de UER Sciences de la Nature

Le Directeur,

Vu et accordé le permis d'imprimer.

Lyon, le 14 avril 1972

Le Président de l'Université
Claude Bernard,

J. BOIDIN

